Journal of Machine Learning Research 26 (2025) 1-44 Submitted 12/23; Revised 7/25; Published 9/25

Operator Learning for Hyperbolic Partial Differential Equations

Christopher Wang CYW33@CORNELL.EDU
Department of Mathematics
Cornell University
Ithaca, NY 14853, USA

Alex Townsend TOWNSEND@CORNELL.EDU
Department of Mathematics
Cornell University
Ithaca, NY 14853, USA

Editor: Fei Sha

Abstract

We construct the first rigorously justified probabilistic algorithm for recovering the solution operator of a hyperbolic partial differential equation (PDE) in two variables from input-output training pairs. The primary challenge of recovering the solution operator of hyperbolic PDEs is the presence of characteristics, along which the associated Green's function is discontinuous. Therefore, a central component of our algorithm is a rank detection scheme that identifies the approximate location of the characteristics. By combining the randomized singular value decomposition with an adaptive hierarchical partition of the domain, we construct an approximant to the solution operator using $O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ input-output pairs with relative error $O(\Xi_\epsilon^{-1}\epsilon)$ in the operator norm as $\epsilon \to 0$, with high probability. Here, Ψ_ϵ represents the existence of degenerate singular values of the solution operator, and Ξ_ϵ measures the quality of the training data. Our assumptions on the regularity of the coefficients of the hyperbolic PDE are relatively weak given that hyperbolic PDEs do not have the "instantaneous smoothing effect" of elliptic and parabolic PDEs, and our recovery rate improves as the regularity of the coefficients increases. We also include numerical experiments which corroborate our theoretical findings.

Keywords: Data-driven PDE learning, hyperbolic PDE, operator learning, low-rank approximation, randomized SVD

1. Introduction

In this paper, we consider the recovery of the solution operator associated with hyperbolic partial differential equations (PDEs) from data. Generally, the task of PDE learning is to capture information about an unknown, inhomogeneous PDE given data corresponding to the observed effect of the PDE on input functions (Fan et al., 2019; Fan and Ying, 2020; Gin et al., 2021; Karniadakis et al., 2021; Kovachki et al., 2023; Li et al., 2020a,b, 2021; Lu et al., 2021a,b; Wang et al., 2021). These PDEs typically represent real-world dynamical systems whose governing principles are poorly understood, even though one can accurately observe or predict their evolutions, either through experimentation or through simulation. Data-driven PDE learning has applications in climate science (Bi et al., 2023; Lam et al., 2023), biology (Raissi et al., 2020), and physics (Chen and Gu, 2021; Kochkov et al., 2021; Kutz, 2017; Qian et al., 2020), and is a significant area of research in scientific machine learning and reduced order modeling (Berman and Peherstorfer, 2023, 2024; Brunton et al., 2016; Chen et al., 2024; de Hoop et al., 2023; Krishnapriyan et al., 2021; Rudy et al., 2017; Subramanian et al., 2023; Zhang and Lin, 2018). In particular, many effective practical schemes based on neural networks have been developed to recover the solution operator associated with an unknown PDE (Boullé et al., 2022a; Feliu-Faba et al., 2020; Fan and Ying, 2020; Gin et al., 2021; Li et al., 2020a,b, 2021; Wan et al., 2023; Wang et al., 2021), although the inscrutability of these neural networks as "black boxes" often prevents one from understanding the underlying explanation for their success. Moreover, theoretical research in PDE learning typically centers on error estimates for neural network-based schemes (Kovachki et al., 2023; Lanthaler et al., 2022; Lu et al., 2021a). There is a growing body of research for operator learning in the context of elliptic and parabolic PDEs (Boullé et al., 2022b; Boullé and Townsend, 2023; Schäfer and Owhadi, 2024; Schäfer et al., 2021), but there is a lack of theoretical work on solution operators of hyperbolic PDEs. The existing work focuses primarily on deep learning models for solving hyperbolic PDEs (Arora, 2023; Berman and Peherstorfer, 2024; Bruna et al., 2024; Guo et al., 2020; Huang and Agarwal, 2023; Rodriguez-Torrado et al., 2022; Thodi et al., 2022) or related operators associated with inverse problems (Khoo and Ying, 2019; Li et al., 2022; Molinaro et al., 2023), rather than recovering their solution operators. Therefore, we aim to shed light on the theoretical aspects of solution operator learning for hyperbolic PDEs and to open the field for further research.

We consider an unknown second-order hyperbolic linear partial differential operator (PDO) in two variables of the form:

$$\mathcal{L}u := u_{tt} - a(x,t)u_{xx} + b(x,t)u_x + c(x,t)u, \qquad (x,t) \in D_T := [0,1]\times[0,1] \tag{1a}$$

together with homogeneous initial-boundary conditions

$$\left\{\begin{aligned} u(x,0) = u_t(x,0) = 0, &\qquad 0 \le x \le 1 \\ u(0,t) = u_x(0,t) = 0, &\qquad 0 \le t \le 1 \\ u(1,t) = u_x(1,t) = 0, &\qquad 0 \le t \le 1 \end{aligned}\right\}. \tag{1b}$$

We assume that the coefficients are somewhat regular, namely, $a, b, c \in \mathcal{C}^1(D_T)$, and that $\mathcal{L}$ is strictly hyperbolic, i.e., $a > 0$, and self-adjoint, i.e., $a_x + b = 0$.[1] For equations of the form $\mathcal{L}u = f$, the function f is called the forcing term of the PDE, while u is the corresponding system's response or solution. The hyperbolic equation $\mathcal{L}u = f$ describes wave-like phenomena, especially in heterogeneous media, such as water waves, acoustic and electromagnetic signals, or earthquakes (Evans, 2010; Lax, 2006).

Our learning task is to approximate the solution operator that maps forcing terms to responses, given training data in the form of input-output pairs $\{(f_j, u_j)\}_{j=1}^N$ that satisfy either the Cauchy problem (1) or the corresponding adjoint Cauchy problem (see Section 4 for details). Associated with the Cauchy problem (1) is a unique Green's function $G : D_T \times D_T \to \mathbb{R}$, which is a kernel for the solution operator (Courant and Hilbert, 1962;

[1] We discuss relaxations of these constraints, as well as inhomogeneous initial conditions, in Section 6.

Mackie, 1965). That is, solutions to the problem (1) are given by the integral operator

$$u(x,t) = \int_{D_T} G(x,t;y,s)f(y,s)\,\mathrm{d}y\,\mathrm{d}s, \qquad (x,t) \in D_T, \tag{2}$$

for $f \in L^2(D_T)$. We call G the *homogeneous Green's function.* Our goal is to recover the action of the integral operator in (2) as accurately as possible, measured by the operator norm. Notably, *we are not solving an inverse problem*, in that we are not interested in learning the coefficients of (1a).[2] In fact, due to the compactness of the solution operator, its recovery is well-posed, and our proposed algorithm is stable (see Section A.1). Rather, our task is to construct a direct solver for the Cauchy problem (1) without explicit knowledge of the equation's coefficients.

1.1 Challenges and contributions

This paper describes the first theoretically justified scheme for recovering the solution operator associated with hyperbolic PDEs using only input-output data. We also provide a rigorous rate of recovery for our algorithm. While we adopt some of the strategies used by Boullé et al. (2022b) and Boullé and Townsend (2023), such as the randomized singular value decomposition (rSVD), we face three unique circumstances when recovering the Green's functions of hyperbolic PDEs that make our situation challenging:

Characteristic curves

The primary difficulty of recovering the Green's function of a hyperbolic PDE is the presence of characteristic curves (or simply characteristics), along which the Green's function is highly irregular. Characteristics describe the trajectory of waves in spacetime and are determined by the coefficients of (1a), so their location is also unknown to us in advance. Thus, our recovery algorithm needs to detect if a region of the Green's function's domain intersects a characteristic curve, using only input-output data.

Adaptive partitioning

For elliptic and parabolic PDEs, the Green's functions are numerically low-rank off the diagonal, so they are well-suited to an approximation strategy via hierarchical partitioning of the domain (Bebendorf and Hackbusch, 2003; Boullé et al., 2022b). Since the Green's functions of hyperbolic PDEs have numerically high rank not only on the diagonal but also along the characteristics, we cannot apply a naive hierarchical partition of the domain. Instead, we use an adaptive partitioning strategy, which at each level uses information from the rSVD to decide which regions to partition further.

Regularity of coefficients

Underlying the difficulty posed by characteristic curves is a more fundamental issue with hyperbolic PDEs: irregularities are not instantaneously smoothed and propagate. This feature of Green's functions makes them challenging to recover, as the variable coefficients can

[2] Learning the coefficients of an equation can be done with techniques similar to sparse identification of nonlinear dynamics (SINDy) (Brunton et al., 2016; Kaiser et al., 2018).

create additional discontinuities. Accordingly, the recovery rate we derive for the hyperbolic case depends on the regularity of the coefficients; the more regular the coefficients, the faster the recovery.

To overcome these challenges, we rely on the following three properties of the Green's function G of the hyperbolic PDO of (1a) (see Section 2):

1. The function G is square-integrable, with jump discontinuities along the characteristics and on the diagonal of $D_T \times D_T$.
2. Away from the diagonal and the characteristics, G is regular and, consequently, numerically low-rank.
3. The characteristics form a piecewise regular hypersurface in $D_T \times D_T$ and never "accumulate" in finite time. In other words, the volume of a tube of radius δ around the characteristic surface shrinks to zero as $\delta \to 0$.

Our main technical contribution is the derivation of a rigorous probabilistic algorithm that constructs an approximant to the solution operator associated with (1a) using randomly generated input-output data $\{(f_j, u_j)\}_{j=1}^{N}$, with a small error in the operator norm. In particular, we show in Theorem 6 that with high probability, the solution operator can be recovered within a tolerance of $\Xi_\epsilon^{-1}\epsilon$ using $O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ input-output pairs, where Ξ_ϵ and Ψ_ϵ are defined in (31) and (32) and describe features of the operator $\mathcal{L}$ and the input-output data—namely, the size of the singular value gaps of the solution operator, and the quality of the covariance kernel used to generate the training data—that cannot be controlled without additional assumptions. Our construction relies on an adaptive hierarchical partition of the spatio-temporal domain, which roughly identifies the location of the characteristic curves, combined with the rSVD for HS operators. We remark that the learning rate is comparable to the one derived by Boullé and Townsend (2023) for elliptic PDEs, where it was later shown that an ϵ^{-6} factor can be improved to polylog$(1/\epsilon)$ using the peeling algorithm; see Section 6.2 (Boullé et al., 2023; Levitt and Martinsson, 2024; Lin et al., 2011). We also emphasize that our scheme succeeds for equations with *both space- and time-dependent* coefficients, which informs the use of the time domain rather than the frequency domain in our analysis. While we are interested in as general of a setting as possible, other schemes that exploit the Helmholtz equation through the frequency domain may be preferable when the coefficients are time-independent (Anderson et al., 2020; Liu et al., 2023; Zepeda-Núñez and Demanet, 2016).

A key component of our probabilistic algorithm is a scheme that detects the numerical rank of an operator using input-output data, which allows us to tell whether or not a domain intersects a characteristic. We do so by showing, in Theorem 5, that the rSVD can efficiently recover an operator's dominant singular subspaces. We then use the singular subspaces to approximate the dominant singular values of the operator, whose decay rate corresponds to the operator's numerical rank. While our scheme assumes the existence of a gap between adjacent singular values, such an assumption is reasonable in practice since the singular values of the solution operator typically exhibit decay, which is fast enough for efficient numerical rank detection (Meier and Nakatsukasa, 2024). Still, we believe that the assumption of a singular value gap can be discarded, although this will not be discussed in the present work.

Our rank detection scheme facilitates the adaptive feature of the partitioning strategy, since at each hierarchical level we partition only the subdomains flagged as numerically high-rank. While such adaptive strategies are not new and have been previously applied to wave-like settings (Liu et al., 2021; Massei et al., 2022; Zepeda-Núñez and Demanet, 2016), our work supplies, to our knowledge, the first theoretical guarantee that the rSVD can be used in an adaptive scheme in a stable manner and with high probability of success. The use of the rSVD is particularly important in the realm of operator learning, where one often only has access to input-output data and thus cannot compute a SVD directly.

Finally, in Theorem 1, we improve the error estimates for the rSVD for HS operators, as derived by Boullé and Townsend (2023), by a factor of $\sqrt{k}$, where k is the target rank of the constructed approximant. Assuming one always oversamples using an additional k training pairs, then the error factor in Boullé and Townsend (2023, Thm. 1) grows to infinity roughly like $O(k)$ as $k \to \infty$, whereas, practically speaking, our error factor remains bounded. This improvement shows that the rSVD for HS operators behaves similarly to the rSVD for matrices; our bounds are asymptotically comparable to the bounds proved in Halko et al. (2011).

We remark that because our algorithm is mainly of theoretical value, we choose to work in the continuous setting—that is, without considering discretization—both to simplify the analysis and to maintain discretization-invariance of our theoretical guarantees. It may be desirable to learn the Green's function as an operator between infinite-dimensional function spaces rather than to learn a space- and time-discretized version of the Green's function. Indeed, Huang et al. (2025) shows that learning the solution operator as a "function-to-function" map and only discretizing its inputs and outputs when necessary may be more data-efficient than learning its discretization as a "vector-to-vector" map. In practice, learning a solution operator without discretizing in space or time can be achieved by working in a finite-dimensional subspace of L^2 using, for instance, the Legendre basis to represent functions, as done in the MATLAB package `chebfun` (Driscoll et al., 2014). Nevertheless, we implement a space- and time-discretized version of our algorithm in Section 5, demonstrating its robustness to discretization.

1.2 Organization

The paper proceeds as follows. In Section 2, we review the characteristics of hyperbolic PDEs and their relationship with the Green's function, while Section 3 develops the necessary tools from randomized linear algebra, namely, the rSVD for Hilbert–Schmidt (HS) operators. These tools are employed in Section 4 to construct our probabilistic algorithm for recovering the solution of a hyperbolic PDO using input-output training pairs; we also analyze the recovery rate and probability of success. A numerical implementation and example of our algorithm is presented in Section 5. Finally, we summarize our results and discuss further directions of research in Section 6. Background material on HS operators, quasimatrices, orthogonal projectors, Gaussian processes (GPs), and the Legendre basis can be found in Appendix A. Proofs related to the rSVD appear in Appendix B.

1.3 Notation

Throughout the paper, we use the following notation. Norms are denoted by $\|\cdot\|$, and the type of norm is given by a subscript. If the argument is an operator, then $\|\cdot\|$ without a subscript denotes the operator norm. For a matrix or quasimatrix $\mathbf{A}$ (see Section A.2), we denote its Moore–Penrose pseudoinverse by $\mathbf{A}^\dagger$. We write $\mathbf{I}_m$ to denote an $m \times m$ identity matrix. For a random object $\mathbf{X}$, we write $\mathbf{X} \sim \mathcal{D}$ to mean that $\mathbf{X}$ is drawn from the distribution $\mathcal{D}$; we write $\mathbf{X} \sim \mathbf{Y}$ to mean that $\mathbf{X}$ has the same distribution as a different random object $\mathbf{Y}$. We use $\mathcal{N}(\mu, \sigma^2)$ to denote the univariate Gaussian distribution with mean μ and variance σ^2. For a vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and a symmetric positive definite $n \times n$ matrix $\mathbf{C}$, we write $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ for the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. For an integer $r \geq 0$, $\mathcal{C}^r(D)$ denotes the space of r-times continuously differentiable functions on a domain D; if D is closed, then the functions in $\mathcal{C}^r(D)$ are also required to extend continuously to the boundary of D. The other function space we use is the Hilbert space $L^2(D)$ of square-integrable functions on D.

2. Green's functions of hyperbolic PDOs

Our recovery scheme relies crucially on understanding where the singularities of the Green's function lie. In this section, we investigate the geometry of characteristic curves and discuss the dependence of the regularity of the Green's function on the regularity of the coefficients of $\mathcal{L}$ as in (1a). We assume that the coefficients of $\mathcal{L}$ have regularity $a, b, c \in \mathcal{C}^r(D_T)$, for integer $r \geq 1$.

2.1 Green's functions in the domain $\mathbb{H}$

We first consider the homogeneous Green's function $G^{\mathbb{H}}$ of $\mathcal{L}$, as defined in (2), in the unbounded domain $\mathbb{H} = \mathbb{R} \times [0, \infty)$.[3] Since we assume the coefficients a, b, c are at least continuously differentiable, G exists and is unique (Courant and Hilbert, 1962, Ch. V.5–6). Its discontinuities lie on the characteristics, which we now describe.

The characteristic curves passing through some point $(x_0, t_0) \in \mathbb{H}$ can be interpreted as the paths in spacetime traversed by waves propagating from an instantaneous unit force at (x_0, t_0). We obtain a family of characteristic curves by iterating over all $(x_0, t_0) \in \mathbb{H}$. For hyperbolic equations in two variables, the characteristic curves are the graphs of solutions to the following ordinary differential equations:

$$x'(t) + \sqrt{a(x(t), t)} = 0, \qquad x'(t) - \sqrt{a(x(t), t)} = 0 \tag{3}$$

over all choices of initial conditions (Courant and Hilbert, 1962, Ch. III.1). Thus, for any $(x_0, t_0) \in \mathbb{H}$, there exist exactly two characteristic curves passing through (x_0, t_0), given by the equations in (3), both of which are of class $\mathcal{C}^{r+1}$ and satisfy the initial condition $x(t_0) = x_0$. Viewed as functions of t, one is strictly increasing, and the other is strictly

[3] Much of the classical literature discusses what is known as the *Riemann function*, rather than the Green's function, of hyperbolic equations. The Riemann function (also referred to as the Riemann–Green function or radiation solution) is a fundamental solution of the PDE and allows one to write an integral solution to the homogeneous equation given Cauchy initial data. Essentially, every result for the Riemann function also holds for the homogeneous Green's function, as they are closely related (Mackie, 1965).

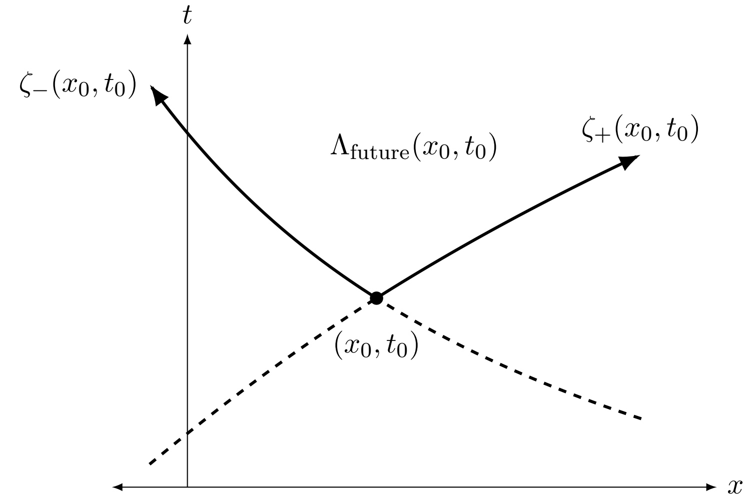

Figure 1: Characteristics associated with the coefficient $a(x,t) = \frac{(x+1)^2+1}{t+1}$ in the domain $\mathbb{H}$, with initial point $(x_0,t_0) = (1,1)$. They represent wave trajectories in spacetime produced by a unit force at $(1,1)$: solid curves are future trajectories, while dashed curves are past trajectories. Positive and negative characteristic rays emanating from (x_0,t_0) are solid and labeled by $\zeta_\pm(x_0,t_0)$. The future light cone is the region labeled Λ_{future}.

decreasing, so they intersect transversally at (x_0,t_0) and partition $\mathbb{H}$ into four connected components lying respectively to the north, south, east, and west of (x_0,t_0). We refer to the component to the north—that is, forward in time—as the *future light cone* $\Lambda_{\text{future}}(x_0,t_0)$. Additionally, we call the characteristic "ray" emanating toward the northwest of (x_0,t_0) the *negative characteristic ray*, and likewise we call the ray emanating toward the northeast the *positive characteristic ray* (see Figure 1).

We are interested in the "bundle" of positive and negative characteristic rays indexed over all initial points $(x_0,t_0) \in \mathbb{H}$, since they determine where the Green's function is irregular. Let $\zeta_\pm(x_0,t_0)$ denote the positive and negative characteristic rays, respectively, emanating from (x_0,t_0). Then we define

$$Z_\pm^{\mathbb{H}} := \{(x,t,x_0,t_0) \in \mathbb{H} \times \mathbb{H} : (x,t) \in \zeta_\pm(x_0,t_0)\} \tag{4}$$

as well as

$$Z^{\mathbb{H}} := Z_+^{\mathbb{H}} \cup Z_-^{\mathbb{H}}. \tag{5}$$

Observe that $Z_\pm^{\mathbb{H}}$ are 3-dimensional $\mathcal{C}^r$-manifolds with boundary in $\mathbb{H} \times \mathbb{H}$.[4]

For fixed $(x_0,t_0) \in D_T$, the slices $G_{x_0,t_0}^{\mathbb{H}}$ given by $(x,t) \mapsto G^{\mathbb{H}}(x,t;x_0,t_0)$ are r-times continuously differentiable as long as (x,t) does not lie on either of the characteristic rays

[4] Here, a $\mathcal{C}^r$-manifold is simply a manifold whose transition maps are of class $\mathcal{C}^r$. The claim here follows from the observation that $Z_\pm^{\mathbb{H}}$ can be defined as the graph of the flow generated by the time-dependent vector field $(x,t) \mapsto \sqrt{a(x,t)}$. The details are irrelevant here, and we omit them.

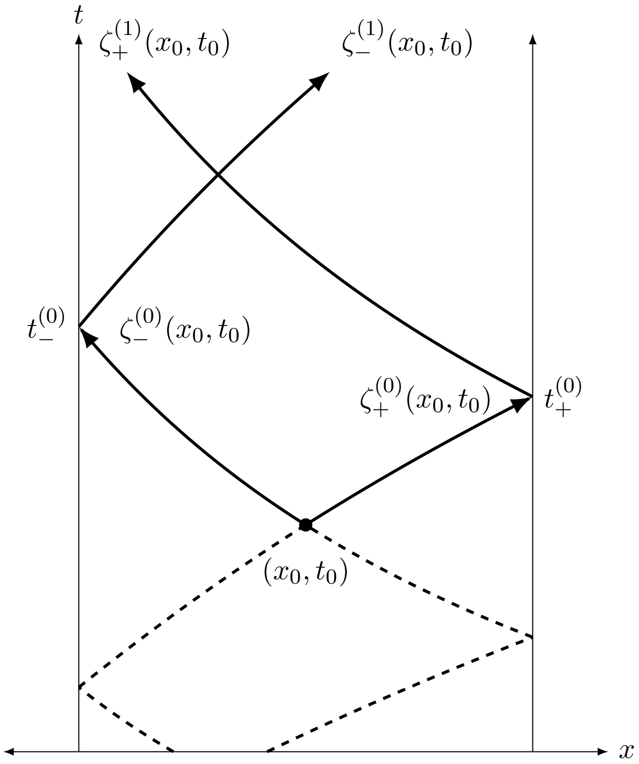

Figure 2: In a bounded domain, the characteristics reflect off the boundary, producing a series of "reflecting characteristic segments," depicted by solid curves labeled $\zeta_\pm^{(j)}(x_0,t_0)$, for $j=0,1,2,\ldots$. "Collision points" are labeled $t_\pm^{(j)}$.

emanating from (x_0,t_0) (Lerner, 1991, Thm. 1). Moreover, due to symmetries of fundamental solutions of hyperbolic equations (Courant and Hilbert, 1962, Ch. V.5), we conclude that $G^{\mathbb{H}}$ has the same regularity in every variable.

For hyperbolic equations in two variables, the singularity of the Green's function on the characteristics is a jump discontinuity. This is because $G^{\mathbb{H}}_{x_0,t_0}$ restricted to $\Lambda_{\text{future}}(x_0,t_0)$ satisfies continuous boundary conditions on the characteristic rays (Courant and Hilbert, 1962, Ch. V.5). Outside the future light cone, the Green's function is identically zero (Mackie, 1965). In summary, we have

$$G^{\mathbb{H}} \in \mathcal{C}^r((\mathbb{H}\times\mathbb{H})\setminus Z^{\mathbb{H}}), \tag{6}$$

where $Z^{\mathbb{H}}$ is defined by (5).

2.2 Green's functions in the domain D_T

When we consider $\mathcal{L}$ on the bounded domain $D_T=[0,1]\times[0,1]$, we must also establish homogeneous boundary conditions on $\{0,1\}\times[0,1]$, in addition to homogeneous initial conditions. Homogeneous boundary conditions produce wave reflections, so the characteristic curves "reflect" off the boundaries. When a positive characteristic ray collides with the right boundary, its reflection is given by the negative characteristic ray emanating from the collision point. Likewise, when a negative characteristic ray collides with the left boundary, its reflection is given by the positive characteristic ray emanating from the collision point.

Wave trajectories can thus be described as a series $\zeta_\pm^{(0)}(x_0,t_0), \zeta_\pm^{(1)}(x_0,t_0), \ldots$ of "reflecting characteristic segments" emanating from (x_0,t_0), which terminates at some index $N(x_0,t_0)$ once the segments cross the time horizon $t=1$ (see Figure 2). Letting $N = \max_{(x_0,t_0)\in D_T} N(x_0,t_0)$, we define, analogous with (4) and (5), the objects

$$Z_\pm^{(j)} := \{(x,t,x_0,t_0) \in D_T \times D_T : (x,t) \in \zeta_\pm^{(j)}(x_0,t_0)\}, \qquad 0 \le j \le N, \tag{7}$$

as well as

$$Z := \bigcup_{j=0}^{N} \left(Z_+^{(j)} \cup Z_-^{(j)} \right). \tag{8}$$

Again, each component $Z_\pm^{(j)}$, $j = 1, \ldots, N$ is a 3-dimensional $\mathcal{C}^r$-manifold with boundary in $D_T \times D_T$, so that Z is a piecewise $\mathcal{C}^r$-manifold. Notice that the number of reflections N is bounded by $\max_{(x,t)\in D_T} \sqrt{a(x,t)}$.[5]

Besides the reflecting characteristics, the homogeneous Green's function G in the domain D_T has the same regularity properties as the homogeneous Green's function $G^{\mathbb{H}}$ in the domain $\mathbb{H}$. The reflections produced by homogeneous boundary conditions manifest as additional jump discontinuities for G, lying on the reflecting characteristic segments described in Section 2.2. In other words, on a bounded domain, we have

$$G \in \mathcal{C}^r((D_T \times D_T) \setminus Z), \tag{9}$$

where Z is defined by (8).

Example 1 *Consider the constant-coefficient wave equation $u_{tt} - a^2 u_{xx} = f$. The Green's function is quite simple to understand and can be written explicitly, but it would be very notationally complicated due to the boundary conditions. It is piecewise constant, and its value in each component is either 0 or $\pm\frac{1}{2a}$ (see Figure 3). Its components are partitioned by characteristic curves, which are lines of slope $\frac{1}{a}$ emanating from the initial points (x_0,t_0) and reflecting off the boundary. For general hyperbolic PDEs of the form (1), the Green's function is usually not piecewise constant; nevertheless, it has the same qualitative properties, in particular jump discontinuities on the characteristics.*

3. Randomized SVD for Hilbert–Schmidt operators

This section introduces our primary tool, which we have adapted from randomized numerical linear algebra. The landmark result of Halko et al. (2011) proved that one could recover the column space of an unknown matrix with high accuracy and a high probability of success by multiplying it with standard Gaussian vectors. Boullé and Townsend (2023) extend this result to HS operators and functions drawn from a non-standard Gaussian process. This algorithm is called the rSVD.

[5] The claims made in this section can be seen by applying the method of reflections to solve the homogeneous initial-boundary problem which defines the homogeneous Green's function in a bounded domain (see, e.g., Laurent et al., 2021). Again, the details are not relevant to us.

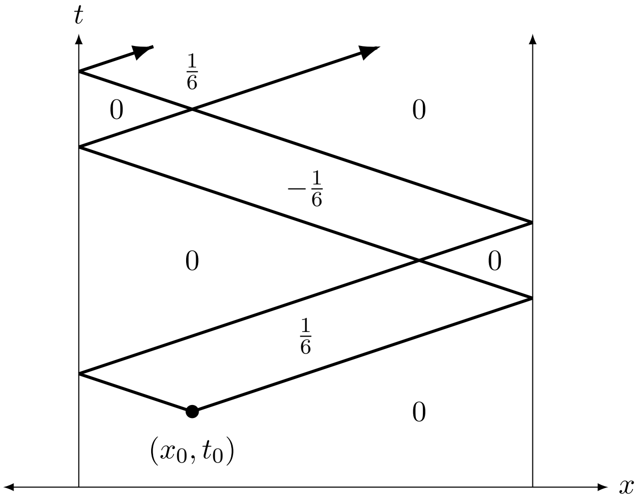

Figure 3: A slice of the Green's function associated with the wave operator $\mathcal{L}u = u_{tt} - 9u_{xx}$ with initial point $(x_0, t_0) = (\frac{1}{4}, \frac{1}{6})$. In this case, the Green's function is piecewise constant with jump discontinuities on the characteristics. The values of the Green's function are labeled in their respective regions.

3.1 Randomized SVD

Given a HS operator $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ with SVE as in (33), we define two quasimatrices $\mathbf{U}$ and $\mathbf{V}$ containing the left and right singular functions of $\mathscr{F}$, so that the jth column of $\mathbf{U}$ and $\mathbf{V}$ is respectively e_j and v_j. We also denote by $\mathbf{\Sigma}$ the infinite diagonal matrix with the singular values $\sigma_1 \geq \sigma_2 \geq \cdots$ of $\mathscr{F}$ on the diagonal. For a fixed integer $k \geq 0$, we define $\mathbf{V}_1$ as the $D_1 \times k$ quasimatrix whose columns are the first k right singular functions $v_1, \ldots, v_k$, and $\mathbf{V}_2$ as the $D_1 \times \infty$ quasimatrix whose columns are $v_{k+1}, v_{k+2}, \ldots$. We analogously define $\mathbf{U}_1$ and $\mathbf{U}_2$ with the left singular functions. Similarly, we define $\mathbf{\Sigma}_1$ as the $k \times k$ diagonal matrix with the first k singular values $\sigma_1, \ldots, \sigma_k$ on the diagonal, and $\mathbf{\Sigma}_2$ as the $\infty \times \infty$ diagonal matrix with $\sigma_{k+1}, \sigma_{k+2}, \ldots$ on the diagonal. In summary, we write

$$\mathscr{F} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{array}{c} \begin{array}{cc} k & \infty \end{array} \\ \begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2\end{bmatrix} \end{array} \begin{array}{cc} \begin{bmatrix}\mathbf{V}_1^* \\ \mathbf{V}_2^*\end{bmatrix} & \begin{array}{c} k \\ \infty \end{array} \end{array} .$$

Additionally, for some fixed continuous symmetric positive definite kernel $K : D_1 \times D_1 \to \mathbb{R}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, we define the infinite matrix $\mathbf{C}$ by

$$[\mathbf{C}]_{ij} = \int_{D_1 \times D_1} v_i(x) K(x, y) v_j(y) \, \mathrm{d}x \, \mathrm{d}y, \qquad i, j \geq 1. \tag{10}$$

Observe that $\mathbf{C}$ is symmetric and positive definite, and that $\mathrm{Tr}(\mathbf{C}) = \mathrm{Tr}(K) < \infty$, so it is a compact operator (Boullé and Townsend, 2023, Lem. 1 and Eq. (11)). We denote by $\mathbf{C}^{-1}$ the inverse operator of $\mathbf{C}$ on the domain for which it is well-defined. Furthermore, for fixed integer $k \geq 1$, we partition $\mathbf{C}$ into

$$\mathbf{C} = \begin{array}{cc} \begin{array}{cc} k & \infty \end{array} & \\ \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} & \begin{array}{c} k \\ \infty \end{array} \end{array}.$$

Since $\mathbf{C}_{11}$ is also positive definite and thus invertible, we define

$$\gamma_k := \frac{k}{\lambda_1 \operatorname{Tr}(\mathbf{C}_{11}^{-1})}, \qquad \xi_k := \frac{1}{\lambda_1 \|\mathbf{C}_{11}^{-1}\|}, \tag{11}$$

which quantify the quality of the covariance kernel K with respect to $\mathscr{F}$. Indeed, the Courant–Fischer minimax principle implies that the jth largest eigenvalue of $\mathbf{C}$ is bounded by λ_j, since $\mathbf{C}$ is a principal submatrix of $\mathbf{V}^* K \mathbf{V}$. It follows that $0 < \gamma_k, \xi_k \leq 1$, and that the best case scenario occurs when the eigenfunctions of K are the right singular functions of $\mathscr{F}$. In that case, $\mathbf{C}$ is an infinite diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, and $\gamma_k = k/(\sum_{j=1}^k \lambda_1/\lambda_j)$ and $\xi_k = \lambda_k/\lambda_1$ attain their minimal values. One can view ξ_k as the operator norm analogue of γ_k, discussed in more detail in Boullé and Townsend (2023, §3.4).

Finally, for some $X \subset D_T$, let $\mathcal{R}_X : L^2(D_T) \to L^2(X)$ denote the restriction to X. Its adjoint $\mathcal{R}_X^* : L^2(X) \to L^2(D_T)$ is the zero extension operator, i.e., $\mathcal{R}_X^* f$ is the function which is equal to f on X and equal to zero everywhere else. We denote by $\mathscr{F}_{X\times Y} := \mathcal{R}_X \mathscr{F} \mathcal{R}_Y^*$ the restriction of $\mathscr{F}$ to $X \times Y$. When considering the restricted operator, we define the analogous quantities

$$\gamma_{k,X\times Y} := \frac{k}{\lambda_1 \operatorname{Tr}(\mathbf{C}_{11,X\times Y}^{-1})}, \qquad \xi_{k,X\times Y} := \frac{1}{\lambda_1 \|\mathbf{C}_{11,X\times Y}^{-1}\|}, \tag{12}$$

where $\mathbf{C}_{11,X\times Y}$ is the $k \times k$ matrix

$$[\mathbf{C}_{11,X\times Y}]_{ij} := \int_{D_1\times D_1} \mathcal{R}_Y^* v_{i,X\times Y}(x) K(x,y) \mathcal{R}_Y^* v_{j,X\times Y}(y) \,\mathrm{d}x \,\mathrm{d}y, \qquad 1 \leq i,j \leq k,$$

and $v_{1,X\times Y}, \ldots, v_{k,X\times Y}$ are the dominant right k-singular functions of $\mathscr{F}_{X\times Y}$. We also define $\sigma_{1,X\times Y} \geq \sigma_{2,X\times Y} \geq \cdots$ as the singular values of $\mathscr{F}_{X\times Y}$.

In this notation, we state an analogue of rSVD for HS operators (Boullé and Townsend, 2023, Thm. 1) with respect to the operator norm. We improve the error bound by a factor of k compared to Boullé and Townsend (2023, Thm. 1).

Theorem 1 *Let $D_1, D_2 \subseteq \mathbb{R}^d$ be domains with $d \geq 1$, and let $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ be a HS operator with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Select a target rank $k \geq 2$, an oversampling parameter $p \geq 2$, and a $D_1 \times (k+p)$ quasimatrix $\mathbf{\Omega}$ such that each column is independently drawn from $\mathcal{GP}(0, K)$, where $K : D_1 \times D_1 \to \mathbb{R}$ is a continuous symmetric positive definite kernel with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Set $\mathbf{Y} = \mathscr{F}\mathbf{\Omega}$. Then*

$$\mathbb{E}\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\| \leq \left(1 + \frac{1}{\xi_k} + \sqrt{\frac{\operatorname{Tr}(K)}{\lambda_1 \xi_k}} \cdot \frac{e\sqrt{k+p}}{p} + \sqrt{\frac{k}{\gamma_k (p+1)}}\right) \sigma_{k+1}, \tag{13}$$

where γ_k, ξ_k are defined in (11)*. Moreover, if $p \geq 4$, then for any $s, t \geq 1$, we have*

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\| \leq \left[1 + \frac{1}{\xi_k} + \frac{e}{\sqrt{\xi_k}}\left(s + \sqrt{\frac{\mathrm{Tr}(K)}{\lambda_1}}\right)\frac{\sqrt{k+p}}{p+1} \cdot t + \sqrt{\frac{k}{\gamma_k(p+1)}} \cdot t\right]\sigma_{k+1} \tag{14}$$

with probability $\geq 1 - 2t^{-p} - e^{-s^2/2}$.

Proof See Appendix B. ■

We also present a simplified version of the probability bound in (14).

Corollary 2 *Under the assumptions of Theorem 1 with $k = p \geq 4$, we have*

$$\|\mathscr{F} - \mathbf{P}_{\mathbf{Y}}\mathscr{F}\| \leq \left[1 + \frac{1}{\xi_k}\left(19 + 11\sqrt{\frac{\mathrm{Tr}(K)}{\lambda_1 k}}\right)\right]\sigma_{k+1} \tag{15}$$

with probability $\geq 1 - 3e^{-k}$.

Proof We evaluate (14) by selecting $s = \sqrt{2k}$ and $t = e$. We also use the inequalities

$$\frac{1}{\sqrt{\gamma_k}} \leq \frac{1}{\sqrt{\xi_k}} \leq \frac{1}{\xi_k},$$

which follow from the fact that $0 < \gamma_k, \xi_k \leq 1$ and $\mathrm{Tr}(\mathbf{C}_{11}^{-1}) \leq k\|\mathbf{C}_{11}^{-1}\|$. ■

For convenience, we abbreviate the error factors in (14) and (15) by

$$A_{k,p}(s,t) := 1 + \frac{1}{\xi_k} + \frac{e}{\sqrt{\xi_k}}\left(s + \sqrt{\frac{\mathrm{Tr}(K)}{\lambda_1}}\right)\frac{\sqrt{k+p}}{p+1} \cdot t + \sqrt{\frac{k}{\gamma_k(p+1)}} \cdot t, \tag{16}$$

$$A_k := 1 + \frac{1}{\xi_k}\left(19 + 11\sqrt{\frac{\mathrm{Tr}(K)}{\lambda_1 k}}\right). \tag{17}$$

as well as the analogue

$$A_{k,X\times Y} = 1 + \frac{1}{\xi_{k,X\times Y}}\left(19 + 11\sqrt{\frac{\mathrm{Tr}(K)}{\lambda_1 k}}\right) \tag{18}$$

when considering a restricted domain $X \times Y \subset D_T \times D_T$.

A “power scheme” version of rSVD, as described in Halko et al. (2011) and Rokhlin et al. (2009), allows one to drive down the multiplicative factor in the probabilistic estimate (14) at the expense of a logarithmic factor increase in the number of operator-function products. The idea of the power scheme is that repeated projections of $\mathbf{Y} = \mathscr{F}\mathbf{\Omega}$ onto the column space of $\mathscr{F}$ improves the approximation.

Theorem 3 *Under the same assumptions as Theorem 1, select an integer $q \geq 0$. Let $\mathscr{H} = (\mathscr{F}\mathscr{F}^*)^q\mathscr{F}$ and set $\mathbf{Z} = \mathscr{H}\mathbf{\Omega}$. If $p \geq 4$, then*

$$\|\mathscr{F} - \mathbf{P_Z}\mathscr{F}\| \leq A_{k,p}(s,t)^{1/(2q+1)}\sigma_{k+1} \tag{19}$$

with probability $\geq 1 - 2t^{-p} - e^{-s^2/2}$, where $A_{k,p}(s,t)$ is defined in (16)*.*

Proof See Appendix B. ■

We summarize how the rSVD generates an approximant for the operator $\mathscr{F}$ in Algorithm 1, following Halko et al. (2011).

Algorithm 1 Approximating $\mathscr{F}$ via rSVD

```
Input: HS operator $\mathscr{F}$, GP covariance kernel $K$, target rank $k \geq 4$, oversampling parameter $p \geq 2$, exponent $q \geq 0$, additional parameters $s, t \geq 1$
Output: Approximation $\tilde{\mathscr{F}}$ of $\mathscr{F}$ within relative error given by (19)
Draw a $D_T \times (k+p)$ random quasimatrix $\mathbf{\Omega}$ with independent columns from $\mathcal{GP}(0, K)$
    ▷ See Section A.4 on how to draw such a quasimatrix
Construct $\mathbf{Z} = (\mathscr{F}\mathscr{F}^*)^q\mathscr{F}\mathbf{\Omega}$ by multiplying with $\mathscr{F}$ and $\mathscr{F}^*$
Compute the projector $\mathbf{P_Z}$ via a QR factorization of $\mathbf{Z}$
    ▷ See Trefethen (2010) for Householder triangularization on quasimatrices
Form $\tilde{\mathscr{F}} = \mathbf{P_Z}\mathscr{F}$
```

Remark 4 (Noisy training data) *Here, we quantify the quality of the training data via the terms γ_k and ξ_k, defined in* (11)*, which measure the deviation of the eigenspaces of the chosen covariance kernel K from the dominant right singular subspaces of $\mathscr{F}$. One can also consider training data quality in the form of noise—namely, the presence of random additive perturbation errors arising from the computation of operator-function products, or from the collection of input-output data. In fact, it was proven in the finite-dimensional setting that the rSVD is stable with respect to noise in the input-output data (Boullé et al., 2023, Supp. Info., §2.B), so we assume noiseless data for simplicity.*

3.2 Approximating singular values via singular subspaces

Using the power scheme, we can not only improve the approximation of the operator $\mathscr{F}$ itself but also approximate its singular subspaces. Consequently, we can obtain excellent estimates for the singular values, which in conjunction with Theorem 1 tells us about the numerical rank of $\mathscr{F}$. Notice that while Weyl's theorem (Stewart and Sun, 1990, Cor. 4.10) gives a straightforward bound on an operator's singular values, applying it meaningfully requires those singular values to decay quickly. Conversely, Theorem 5 says that we can approximate the singular values regardless of their decay rate at the cost of additional operator-function products. We emphasize that our argument assumes the existence of a gap between adjacent singular values, although we believe this assumption is not necessary in principle.

Theorem 5 *Assume the hypotheses of Theorem 1, but with $\sigma_k > \sigma_{k+1}$. Select an integer $q \geq 0$ and set δ_q as in (44). Let $\mathscr{H} := (\mathscr{F}\mathscr{F}^*)^q\mathscr{F}$, $\mathbf{Z} := \mathscr{H}\mathbf{\Omega}$, and $\tilde{\mathscr{H}} := \mathbf{P}_{\mathbf{Z}}\mathscr{H}$. Let $\tilde{\mathbf{U}}_k$ be the $D_1 \times k$ quasimatrix whose columns are dominant left k-singular vectors of $\tilde{\mathscr{H}}$. If $\delta_q A_{k,p}(s,t) < 1$, then*

$$\max_{1\leq j\leq k} |\sigma_j - \sigma_j(\tilde{\mathbf{U}}_k^*\mathscr{F})| \leq \frac{2\delta_q A_{k,p}(s,t)}{1-\delta_q A_{k,p}(s,t)}\|\mathscr{F}\| \tag{20}$$

with probability $\geq 1 - 2t^{-p} + e^{-s^2/2}$.

Proof See Appendix B. ∎

4. Recovering the Green's function

In this section, we construct a global approximant of the homogeneous Green's function of a 2-variable hyperbolic PDO $\mathcal{L}$ of the form (1a) in the domain $D_T = [0,1]\times[0,1]$. Recall that $\mathcal{L}$ is assumed to be linear strictly hyperbolic, and self-adjoint, with coefficients satisfying $a, b, c \in \mathcal{C}^1(D_T)$. We suppose that one can generate N forcing terms $\{f_j\}_{j=1}^N$ drawn from a Gaussian process with continuous symmetric positive definite covariance kernel $K : D_T \times D_T \to \mathbb{R}$ and use them to query the associated solution operator $\mathscr{F}$ of $\mathcal{L}$, as well as its adjoint $\mathscr{F}^*$, to generate solutions $u_j = \mathscr{F}f_j$ or $u_j = \mathscr{F}^*f_j$.[6] We derive a bound on the number of input-output pairs $\{(f_j, u_j)\}_{j=1}^N$ needed to approximate the Green's function within a given error tolerance measured in the operator norm, with a high probability of success. Our result is summarized in the following theorem.

Theorem 6 *Let $\mathcal{L}$ be a hyperbolic PDO given in* (1a)*. Let $G : D_T \times D_T \to \mathbb{R}$ be the homogeneous Green's function of $\mathcal{L}$ in the domain D_T, and let $\mathscr{F} : L^2(D_T) \to L^2(D_T)$ be the solution operator with kernel G, as in* (2)*. Additionally, define Ξ_ϵ and Ψ_ϵ as in* (31) *and* (32)*. For any sufficiently small $\epsilon > 0$ such that $\Psi_\epsilon > 0$, there exists a randomized algorithm that can construct an approximation $\tilde{\mathscr{F}}$ of $\mathscr{F}$ using $O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ input-output training pairs of $\mathcal{L}$, such that*

$$\|\mathscr{F} - \tilde{\mathscr{F}}\| = O(\Xi_\epsilon^{-1}\epsilon)\|\mathscr{F}\|$$

with probability $\geq 1 - O(e^{-1/\epsilon})$.

Remark 7 (Increased regularity of coefficients) *The number of input-output training pairs can be reduced by assuming greater regularity of the coefficients a, b, c of $\mathcal{L}$. In particular, if $a, b, c \in \mathcal{C}^r(D_T)$ for some $r \geq 1$, then Theorem 6 easily generalizes, via* (40) *and* (9)*, so that only $O(\Psi_\epsilon^{-1}\epsilon^{-(6+1/r)}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ training pairs are needed to approximate $\mathscr{F}$ with relative error $O(\Xi_\epsilon^{-1}\epsilon)$. If we assume a, b, c are analytic, the $\epsilon^{-1/r}$ factor can be reduced even further to $\log(\epsilon^{-1})$.*

[6] We note that $\mathscr{F}^*$ is the solution operator of the adjoint Cauchy problem of (1), meaning that $u = \mathscr{F}^*f$ satisfies both the adjoint PDE $\mathcal{L}^*u = f$ as well as homogeneous conditions at the boundary and at the *terminal* time $t = 1$. In other words, the adjoint problem is the backward-in-time version of (1). In particular, $\mathscr{F}$ is self-adjoint if and only if $\mathcal{L}$ is self-adjoint and the coefficients a, b, c are time-independent.

A randomized algorithm that achieves Theorem 6 is summarized in Algorithm 2, and we dedicate the remainder of this section to constructing it. We fix a sufficiently small $0 < \epsilon < 1/2$, so that by (40) and (9), there exists $k_\epsilon := \lceil C\epsilon^{-1} \rceil \geq 4$, where C is a constant depending only on $\max_{(D_T \times D_T)\setminus Z} |\nabla G|$ and Z is defined in (8). This ensures that $\text{rank}_\epsilon(\mathscr{F}_{X\times Y}) < k_\epsilon$ holds whenever $X \times Y \cap Z = \varnothing$. Here, Z is the bundle of reflecting characteristic segments, defined in (8). We also set the parameters $s = \sqrt{2k_\epsilon}$, $t = e$, and $p = k_\epsilon$, so that the probabilities of failure for Theorems 1, 3, and 5 are bounded by $3e^{-k_\epsilon}$ (see Corollary 2).

Algorithm 2 Learning the solution operator via input-output data

```
Input: Black-box solver for $\mathcal{L}$, GP covariance kernel $K$, tolerance $0 < \epsilon < 1/2$
Output: Approximation $\tilde{\mathscr{F}}$ of the solution operator $\mathscr{F}$ within relative error $\Xi_\epsilon^{-1}\epsilon$
Set target rank $k_\epsilon \geq 4$
while $\text{vol}(D_{\text{red}}(L)) > \epsilon^2$ do
    for $D \in \mathcal{D}_{\text{red}}(L)$ do
        Partition $D$ into 16 subdomains $D_1, \ldots, D_{16}$                  ▷ Section 4.2
        for $i = 1 : 16$ do
            Determine the numerical rank of $\mathscr{F}$ on $D_i$ (Algorithm 3)    ▷ Section 4.1
            if $\mathscr{F}$ is numerically low-rank on $D_i$ then
                Color $D_i$ green
                Approximate $\mathscr{F}$ on $D_i$ using the rSVD (Algorithm 1)    ▷ Section 4.3
            else
                Color $D_i$ red and add $D_i$ to $\mathcal{D}_{\text{red}}(L+1)$
Pad the approximant $\tilde{\mathscr{F}}$ with zeros on the remaining red subdomains   ▷ Section 4.4.1
```

4.1 Rank detection scheme

Since the locations of the characteristics are unknown, we adaptively partition $D_T \times D_T$ in a hierarchical manner. The adaptive part relies on a "rank detection scheme" that detects the numerical rank of $\mathscr{F}$ in a given subdomain $X \times Y \subset D_T \times D_T$ (see Algorithm 3).

Similar to Boullé and Townsend (2023, §4.1.2), we generate a $Y \times 2k_\epsilon$ quasimatrix $\mathbf{\Omega}$ such that each column is independently drawn from a Gaussian process defined on Y, given by $\mathcal{GP}(0, \mathcal{R}_{Y\times Y}K)$, where $\mathcal{R}_{Y\times Y}$ is the operator that restricts functions to the domain $Y \times Y$. Then we extend by zero each column of $\mathbf{\Omega}$ from $L^2(Y)$ to $L^2(D_T)$ in the form $\mathcal{R}_Y^*\mathbf{\Omega}$.

Next, we select

$$q_{\epsilon,X\times Y} := \max\left(0, \left\lceil \frac{1}{2}\left(\frac{\log(1 + A_{k_\epsilon,X\times Y} + 2A_{k_\epsilon,X\times Y}/\epsilon)}{\log(\sigma_{k_\epsilon,X\times Y}/\sigma_{k_\epsilon+1,X\times Y})} - 1\right)\right\rceil\right), \tag{21}$$

where $A_{k_\epsilon,X\times Y}$ is defined in (18), and set $\mathscr{H} = (\mathscr{F}\mathscr{F}^*)^{q_{\epsilon,X\times Y}}\mathscr{F}$. We approximate the range of $\mathscr{H}$ via $\mathbf{Z} = \mathscr{H}\mathcal{R}_Y^*\mathbf{\Omega}$ and then compute the rank-$2k_\epsilon$ approximant $\tilde{\mathscr{H}}_{X\times Y} = \mathbf{P}_{\mathcal{R}_X\mathbf{Z}}\mathcal{R}_X\mathscr{H}\mathcal{R}_Y^*$. Since $\tilde{\mathscr{H}}_{X\times Y}$ has finite rank, we can compute its SVD and extract the quasimatrix $\tilde{\mathbf{U}}_{k_\epsilon,X\times Y}$ whose columns are the dominant left k_ϵ-singular vectors of $\tilde{\mathscr{H}}_{X\times Y}$, and finally compute the dominant singular values $\hat{\sigma}_{1,X\times Y} \geq \cdots \geq \hat{\sigma}_{k_\epsilon,X\times Y}$ of the operator $\tilde{\mathbf{U}}_{k_\epsilon,X\times Y}^*\mathscr{F}_{X\times Y}$. In total, these computations require $k_\epsilon(8q_{\epsilon,X\times Y} + 5)$ input-output training pairs, and rely on the querying the adjoint of $\mathscr{F}$.

Algorithm 3 Detecting the numerical rank of $\mathscr{F}$ in a subdomain

```
Input: Black-box solver for $\mathcal{L}$, GP covariance kernel $K$, subdomain $X \times Y$, tolerance
    $0 < \epsilon < 1$
Output: Classification of numerical rank of $\mathscr{F}_{X\times Y}$
Set target rank $k_\epsilon \geq 4$
Set exponent $q_{\epsilon,X\times Y} \approx \log(\epsilon^{-1})$
Construct $\tilde{\mathscr{H}}_{X\times Y}$ using the rSVD (Algorithm 1) with exponent parameter 1 applied to
    $\mathscr{H} = (\mathscr{F}\mathscr{F}^*)^{q_{\epsilon,X\times Y}}\mathscr{F}$
Compute a SVD of $\tilde{\mathscr{H}}_{X\times Y}$ to obtain the dominant left $k_\epsilon$-singular functions $\tilde{\mathbf{U}}_{k_\epsilon,X\times Y}$
Construct $\tilde{\mathbf{U}}^*_{k_\epsilon,X\times Y}\mathscr{F}_{X\times Y}$ using the black-box solver for $\mathcal{L}$
Compute the singular values $\hat{\sigma}_{1,X\times Y} \geq \cdots \geq \hat{\sigma}_{k_\epsilon,X\times Y}$ of $\tilde{\mathbf{U}}^*_{k_\epsilon,X\times Y}\mathscr{F}_{X\times Y}$
if $\hat{\sigma}_{k_\epsilon,X\times Y} < 4\epsilon\hat{\sigma}_{1,X\times Y}$ then $\mathrm{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) < k_\epsilon$
else $\mathrm{rank}_{\epsilon}(\mathscr{F}_{X\times Y}) \geq k_\epsilon$
```

Our choice of $q_{\epsilon,X\times Y}$ in (21) is motivated as follows. Given $\delta_{q_{\epsilon,X\times Y}}$ as defined in (44), we have

$$\frac{2\delta_{q_{\epsilon,X\times Y}}A_{k_\epsilon,X\times Y}}{1-\delta_{q_{\epsilon,X\times Y}}A_{k_\epsilon,X\times Y}} \leq \epsilon,$$

which in conjunction with Theorem 5 implies that the inequalities $\hat{\sigma}_{k_\epsilon,X\times Y} \leq \sigma_{k_\epsilon,X\times Y} + \epsilon\sigma_{1,X\times Y}$ and $(1-\epsilon)\sigma_{1,X\times Y} \leq \hat{\sigma}_{1,X\times Y}$ hold with probability $\geq 1-3e^{-k_\epsilon}$. If $\mathrm{rank}_\epsilon(\mathscr{F}_{X\times Y}) < k_\epsilon$, then $\sigma_{k_\epsilon,X\times Y} \leq \epsilon\sigma_{1,X\times Y}$, hence

$$\hat{\sigma}_{k_\epsilon,X\times Y} \leq 2\epsilon\sigma_{1,X\times Y} \leq \frac{2\epsilon}{1-\epsilon}\hat{\sigma}_{1,X\times Y}.$$

Since $\epsilon < 1/2$, then the right-hand side is bounded by $4\epsilon\hat{\sigma}_{1,X\times Y}$. On the other hand, if $\hat{\sigma}_{k_\epsilon,X\times Y} \leq 4\epsilon\hat{\sigma}_{1,X\times Y}$, then Theorem 5 again yields

$$\sigma_{k_\epsilon,X\times Y} \leq \hat{\sigma}_{k_\epsilon,X\times Y} + \epsilon\sigma_{1,X\times Y} \leq 5\epsilon\sigma_{1,X\times Y}$$

hence $\mathrm{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) < k_\epsilon$. In summary, we have shown that

$$\mathrm{rank}_\epsilon(\mathscr{F}_{X\times Y}) < k_\epsilon \implies \hat{\sigma}_{k_\epsilon,X\times Y} \leq 4\epsilon\hat{\sigma}_{1,X\times Y} \implies \mathrm{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) < k_\epsilon \tag{22}$$

holds with probability $\geq 1-3e^{-k_\epsilon}$.

The preceding argument implies that we need only to check the validity of the inequality

$$\hat{\sigma}_{k_\epsilon,X\times Y} < 4\epsilon\hat{\sigma}_{1,X\times Y} \tag{23}$$

to determine with high probability whether or not $\mathscr{F}_{X\times Y}$ has numerical rank bounded by k_ϵ, with low error tolerance. In particular, every subdomain that does not intersect Z has $\mathrm{rank}_\epsilon(\mathscr{F}_{X\times Y}) < k_\epsilon$, and the test passes on all such subdomains. Conversely, even if (23) happens to be satisfied for a subdomain that does intersect Z, then that subdomain still has a low numerical rank with comparatively small error tolerance, namely, 5ϵ.

4.2 Hierarchical partition of domain

We now describe the hierarchical partitioning of the domain $D_T \times D_T = [0,1]^4$, so that $\mathscr{F}$ is numerically low-rank on many subdomains, while the subdomains on which $\mathscr{F}$ is not low-rank have a small total volume. Since the probability of failure is very low, we describe the partition deterministically and relegate the discussion of the probabilistic aspects to Section 4.4.4. In particular, we assume throughout this section that (22) holds deterministically.

Our partitioning strategy proceeds as follows. At level $L = 0$, we consider only one subdomain, which is the entirety of $D_T \times D_T$. On each subdomain at the current partition level, we perform the rank detection procedure described in Section 4.1 to check if $\mathscr{F}$ is numerically low-rank on the subdomain. If it is, we color the subdomain green and approximate $\mathscr{F}$ restricted to the subdomain using the rSVD. Otherwise, we color the subdomain red. The next level of the partition then consists of partitioning each red subdomain into smaller subdomains and repeating the coloring process. Since the characteristics lie only in red subdomains, then at each level of the partition we learn the location of the characteristics with finer detail (see Figure 4).

The hierarchical partition of $D_T \times D_T = [0,1]^4$ for n levels is defined recursively as follows.

- At level $L = 0$, the domain $I_{1,1,1,1} := I_1 \times I_1 \times I_1 \times I_1 = [0,1]^4$ is the root of the partitioning tree.
- At a given level $0 \le L \le n-1$, a node $I_{j_1,j_2,j_3,j_4} := I_{j_1} \times I_{j_2} \times I_{j_3} \times I_{j_4}$ is colored red if (23) fails to hold for $X \times Y = I_{j_1,j_2,j_3,j_4}$. Otherwise, the node I_{j_1,j_2,j_3,j_4} is colored green. Green nodes have no children, whereas red nodes have 2^4 children defined as

  $$\{I_{2j_1+k_1} \times I_{2j_2+k_2} \times I_{2j_3+k_3} \times I_{2j_4+k_4} : k_i \in \{0,1\},\ i = 1,2,3,4\}.$$

  Here, if $I_j = [a,b]$, $0 \le a < b \le 1$, then we define $I_{2j} = [a, \frac{a+b}{2}]$ and $I_{2j+1} = [\frac{a+b}{2}, b]$.

Every subdomain $X \times Y$ in levels $0 \le L \le n-1$ is green and satisfies (23), hence $\mathrm{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) < k_\epsilon$, by (22). At the end, we perform the test (23) for all remaining subdomains at level n. Those for which (23) holds are colored green, while the rest are colored red.

4.3 Local approximation of the Green's function

The hierarchical partition of $D_T \times D_T$ described above tells us where $\mathscr{F}$ is numerically low-rank and where it is not. If a subdomain $X \times Y$ is colored green, that is, a subdomain on which (23) holds, then we can approximate it using the rSVD. Since we have generated $\mathbf{Z} = \mathscr{H}\mathcal{R}_Y^*\mathbf{\Omega}$ in the process of rank detection, where $\mathscr{H} = (\mathscr{F}\mathscr{F}^*)^{q_{\epsilon,X\times Y}}\mathscr{F}$ and $\mathbf{\Omega}$ is a quasimatrix with columns drawn from $\mathcal{GP}(0, \mathcal{R}_{Y\times Y}K)$, then we only require an additional $2k_\epsilon$ input-output training pairs to construct $\tilde{\mathscr{F}}_{X\times Y} := \mathbf{P}_{\mathcal{R}_X\mathbf{Z}}\mathcal{R}_X\mathscr{F}\mathcal{R}_Y^*$.[7] By Theorem 3 and (22), we have

$$\|\mathscr{F}_{X\times Y} - \tilde{\mathscr{F}}_{X\times Y}\| \le A_{k_\epsilon,X\times Y}^{1/(2q_{\epsilon,X\times Y}+1)}\sigma_{k_\epsilon+1,X\times Y} \le 5\epsilon A_{k_\epsilon,X\times Y}^{1/(2q_{\epsilon,X\times Y}+1)}\|\mathscr{F}_{X\times Y}\|. \tag{24}$$

[7] In other words, we can skip lines 2–3 in Algorithm 1.

Moreover, this error estimate holds with probability 1, conditioned on the event that (22) is valid (see Section 4.4.4).

On the other hand, if (23) does not hold on $X \times Y$, then we approximate $\mathscr{F}$ by zero. Since the Green's function $G$is continuous in the compact domain $D_T \times D_T$, except for jumps along the characteristics and the diagonal, then there exists a constant C such that

$$G(x,t;x_0,t_0) \leq C\|\mathscr{F}\|, \qquad (x,y,x_0,y_0) \in D_T \times D_T.$$

Hence, for any $X, Y \subset D_T$, we have

$$\|G\|^2_{L^2(X\times Y)} \leq C^2 \operatorname{vol}(X \times Y)\|\mathscr{F}\|^2 \tag{25}$$

and thus

$$\|\mathscr{F}_{X\times Y}\| \leq C\sqrt{\operatorname{vol}(X \times Y)}\|\mathscr{F}\|, \tag{26}$$

where C does not depend on $X \times Y$. Therefore, for sufficiently small domains $X \times Y$, we may approximate $\mathscr{F}_{X\times Y}$ by zero. At the end of partitioning, the total volume of the subdomains on which (23) fails to hold is negligible (see Section 4.4.1).

4.4 Recovering the Green's function on the entire domain

We now show that we can recover $\mathscr{F}$ on the entire domain $D_T \times D_T$.

4.4.1 Global approximation near characteristics

First, we ensure that the volume of all subdomains where $\mathscr{F}$ is not low-rank is small enough to safely ignore that part of $\mathscr{F}$ by approximating it by zero. As one increases the level of hierarchical partitioning, the volume of such subdomains shrinks to zero.

Let $\mathcal{D}_{\text{red}}(L)$ denote the collection of subdomains $X \times Y$ at level L that are colored red, such that $\operatorname{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) \geq k_\epsilon$. Let $D_{\text{red}}(L) := \bigcup_{D\in\mathcal{D}_{\text{red}}(L)} D$. By (9), (22) and the definition of k_ϵ, we have $X \times Y \in \mathcal{D}_{\text{red}}(L)$ only if $(X \times Y) \cap Z \neq \varnothing$. Subdomains at level L are 4-dimensional cubes with side length 2^{-L}, so every $X \times Y \in \mathcal{D}_{\text{red}}(L)$ is contained in the closed tubular neighborhood of Z in $D_T \times D_T$ with radius 2^{-L+1}, defined as

$$T(Z, 2^{-L+1}) := \{p \in D_T \times D_T : \operatorname{dist}(p, Z) \leq 2^{-L+1}\}. \tag{27}$$

The volume of $T(Z, 2^{-L+1})$ depends only on the coefficient a of (1a) and can be computed explicitly using a Weyl-type tube formula.[8] In particular, we have

$$\operatorname{vol}(D_{\text{red}}(L)) \leq \operatorname{vol}(T(Z, 2^{-L+1})) = O(2^{-L}), \tag{28}$$

hence

$$\sum_{X\times Y\in\mathcal{D}_{\text{red}}(L)} \|\mathscr{F}_{X\times Y}\| \leq O(2^{-L/2})\|\mathscr{F}\|$$

by (26). Therefore, by selecting $n_\epsilon := \lceil \log_2(1/\epsilon^2)\rceil$, we guarantee that

$$\sum_{X\times Y\in\mathcal{D}_{\text{red}}(n_\epsilon)} \|\mathscr{F}_{X\times Y}\| = O(\epsilon)\|\mathscr{F}\| \tag{29}$$

and can safely approximate $\mathscr{F}$ by zero on $D_{\text{red}}(n_\epsilon)$.

[8] For $a \in \mathcal{C}^2$, the classical Weyl tube formula suffices, whereas for a with less regularity one requires more complicated expressions; see, e.g., Federer (1959, Thm. 5.6), Gray (2003, Thm. 4.8), and Hug et al. (2004, Eq. (4.4)). We omit the details.

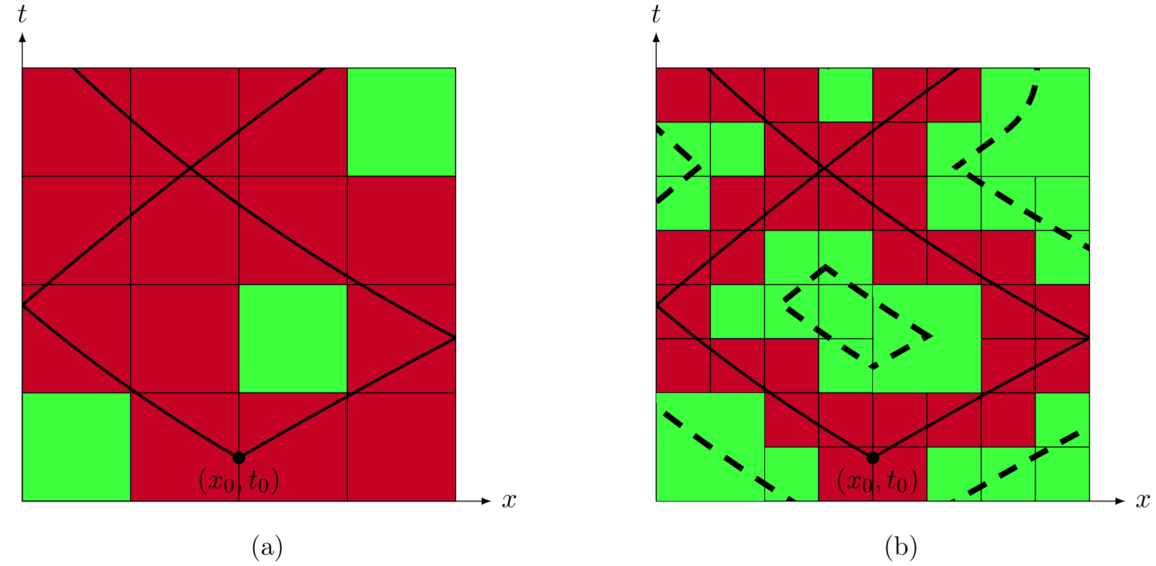

Figure 4: Slices through $D_T \times D_T$ at two levels of hierarchical partitioning. A subdomain intersecting a characteristic is colored red and is further partitioned at the next level. Otherwise, it is colored green. The boundary of the tube of (27) is depicted by the thick dashed lines in (b); notice that every red subdomain is contained within the tube's interior.

4.4.2 Recovery rate

Input-output training pairs are required both for rank detection on subdomains and for approximation for the subdomains on which $\mathscr{F}$ is numerically low-rank. We count the training pairs needed for each task separately.

Since rank detection occurs on every subdomain in the partitioning tree, let $\mathcal{D}_{\text{tree}}(n_\epsilon)$ denote the entire collection of subdomains in the tree after n_ϵ levels of partitioning. For each subdomain colored red at a given level L, we split it into 16 smaller subdomains and perform the rank detection scheme on each, hence

$$\#\mathcal{D}_{\text{tree}}(n_\epsilon) = 16 \sum_{L=0}^{n_\epsilon - 1} \#\mathcal{D}_{\text{red}}(L).$$

Subdomains at the Lth level of partitioning have volume 16^{-L}, so it follows from (28) that

$$\#\mathcal{D}_{\text{red}}(L) \leq 16^L \operatorname{vol}(D_{\text{red}}(L)) = O(8^L).$$

Therefore,

$$\#\mathcal{D}_{\text{tree}}(n_\epsilon) = O(8^{n_\epsilon}) = O(\epsilon^{-6}). \tag{30}$$

Finally, recall from Section 4.1 that performing the rank detection scheme on a subdomain $X \times Y$ requires $k_\epsilon(8q_{\epsilon,X\times Y} + 5)$ training pairs. We remove the dependence on $X \times Y$ by introducing

$$\Xi_\epsilon := \min\{\xi_{k_\epsilon, X\times Y} : X \times Y \in \mathcal{D}_{\text{tree}}(n_\epsilon)\} \in (0, 1], \tag{31}$$

which represents the quality of the covariance kernel K, as well as

$$\Psi_\epsilon := \min\left\{\log\left(\frac{\sigma_{k_\epsilon,X\times Y}}{\sigma_{k_\epsilon+1,X\times Y}}\right) : X\times Y\in\mathcal{D}_{\text{tree}}(n_\epsilon)\right\}, \tag{32}$$

which represents the influence of the singular value gaps in the rank detection scheme. It follows from (18) that $A_{k_\epsilon,X\times Y} = O(\Xi_\epsilon^{-1})$ and thus $q_{\epsilon,X\times Y} = O(\Psi_\epsilon^{-1}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ by (21). Therefore, the total number of training pairs needed for rank detection on every domain in $\mathcal{D}_{\text{tree}}(n_\epsilon)$ is given by

$$N_\epsilon^{(1)} := \sum_{X\times Y\in\mathcal{D}_{\text{tree}}(n_\epsilon)} k_\epsilon(8q_{\epsilon,X\times Y}+5) = O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1})).$$

To count training pairs needed for approximation on domains where $\mathscr{F}$ is numerically low-rank, we first observe that

$$\#\mathcal{D}_{\text{green}}(n_\epsilon) = \sum_{L=1}^{n_\epsilon}\Big(16\cdot\#\mathcal{D}_{\text{red}}(L-1) - \#\mathcal{D}_{\text{red}}(L)\Big) = O(8^{n_\epsilon}),$$

where $\mathcal{D}_{\text{green}}(n_\epsilon)$ is the collection of subdomains $X\times Y$ in the hierarchical level n_ϵ colored green, i.e., those satisfying $\text{rank}_{5\epsilon}(\mathscr{F}_{X\times Y}) < k_\epsilon$. Recall from Section 4.3 that only $2k_\epsilon$ additional training pairs are required to generate an approximant on a subdomain via rSVD after performing rank detection, so the number of training pairs needed for approximation is given by

$$N_\epsilon^{(2)} := \sum_{X\times Y\in\mathcal{D}_{\text{green}}(n_\epsilon)} 2k_\epsilon = O(\epsilon^{-7}).$$

We conclude that the total number of training pairs required is

$$N_\epsilon := N_\epsilon^{(1)} + N_\epsilon^{(2)} = O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1})).$$

4.4.3 Global approximation error

Let $\tilde{\mathscr{F}}$ be the approximant to $\mathscr{F}$ given by stitching together, at the end of the hierarchical partition, the approximants $\tilde{\mathscr{F}}_{X\times Y}$ generated on green subdomains $X\times Y$, and the zero operator on red subdomains. By (29), we have

$$\|\mathscr{F}-\tilde{\mathscr{F}}\| \le \sum_{X\times Y\in\mathcal{D}_{\text{green}}(n)} \|\mathscr{F}_{X\times Y}-\tilde{\mathscr{F}}_{X\times Y}\| + O(\epsilon)\|\mathscr{F}\|,$$

where every term in the summation satisfies (24). Recall that $A_{k_\epsilon,X\times Y} = O(\Xi_\epsilon^{-1})$, hence

$$\|\mathscr{F}-\tilde{\mathscr{F}}\| \le O(\Xi_\epsilon^{-1}\epsilon)\|\mathscr{F}\|$$

is our final error.

4.4.4 Recovery probability

Thus far, we have assumed that our rank detection is guaranteed to succeed on each domain, i.e., that (22) holds deterministically. Circumventing this assumption is simple: since the forcing terms in the input-output training data are independent, and since we perform rank detection on $\#\mathcal{D}_{\text{tree}}(n_\epsilon) = O(\epsilon^{-6})$ domains, we have

$$\mathbb{P}\Big\{(22) \text{ holds for every } X \times Y \in \mathcal{D}_{\text{tree}}(n)\Big\} \geq (1 - 3e^{-k_\epsilon})^{\#\mathcal{D}_{\text{tree}}(n_\epsilon)} = 1 - O(e^{-1/\epsilon}).$$

Notice that for each subdomain, the approximant is generated using the same test quasimatrices as used in the rank detection scheme, hence all error bounds for approximants, e.g., (24), are guaranteed to hold when conditioned on the validity of (22). Therefore, our global probability of failure is $O(e^{-1/\epsilon})$, and we have completed the proof of Theorem 6.

5. Numerical Example

We implement Algorithms 1, 2, and 3 in MATLAB for the constant coefficient wave operator $\mathcal{L}u = u_{tt} - 4u_{xx}$ with homogeneous initial and boundary conditions.[9,10] Our implementation discretizes the domain blockwise at Gauss–Legendre quadrature nodes. Input-output data was generated using the known analytical expression for the true Green's function.

Figure 5 depicts a slice of the approximate Green's function constructed by Algorithm 2 after eight levels of partitioning, as well as the pointwise error in comparison with the actual Green's function. We observe that the rank detection scheme successfully distinguishes between those subdomains that intersect characteristics and those that do not, hence the partition is fine in areas near characteristics and coarse in areas away from characteristics. Errors are concentrated near the characteristics, as expected, while low-rank subdomains are approximated very well. Figure 6 shows the experimental rate of convergence as the amount of input-output training data increases. As predicted by Theorem 6, the error decays like $O(N^{-1/7})$, where N is the number of training data pairs. The slow convergence rate—theoretically predicted and empirically observed—shows that large amounts of data may be needed to accurately approximate Green's functions associated with hyperbolic PDEs. We emphasize that our algorithm is guarding against the nastiest possible solution operators associated with a hyperbolic PDE; if one assumes in advance that the PDE has constant coefficients, then one can do dramatically better. Nevertheless, the numerical example demonstrates that our method is stable and robust to discretization errors.

6. Conclusions and Discussion

Our breakthrough result proves that one can rigorously recover the Green's function of a hyperbolic PDE in two variables using input-output training pairs with high probability.

[9] MATLAB code is available at `https://github.com/chriswang030/OperatorLearningforHPDEs`.

[10] While our code has no difficulty handling more complicated hyperbolic PDEs, the Green's functions for equations with variable coefficients are generally unknown analytically, hence training data must be generated from, say, a numerical solver. Experiments with various popular techniques showed that such solvers were either (a) too slow to be practical, given the large amount of data needed, or (b) introduced numerical dissipation and/or dispersion that dominated the approximation error and resulted in a poor Green's function approximation. We discuss these issues in more detail in Section 6.4.

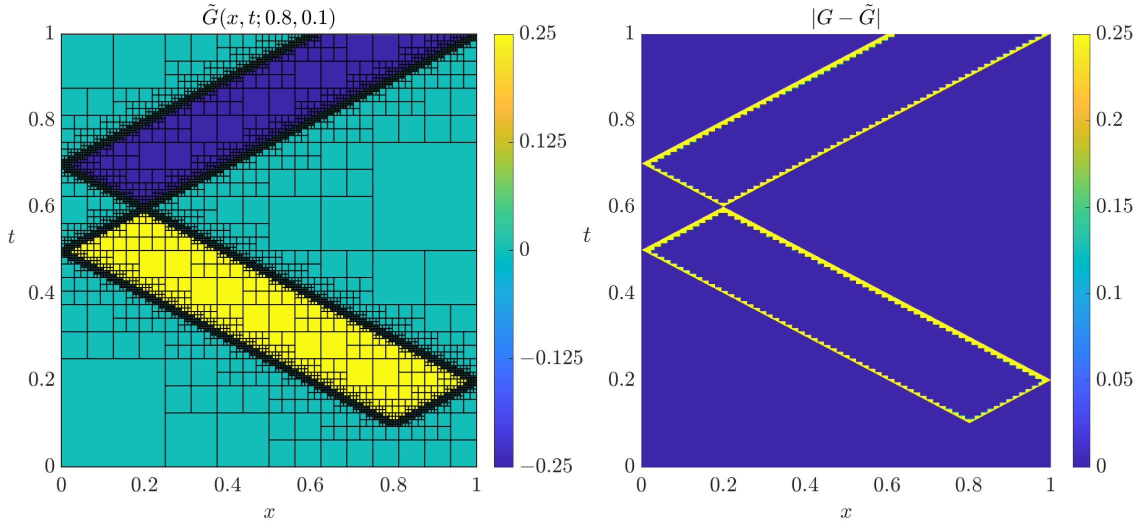

Figure 5: Approximate Green's function at the slice $(y, s) = (0.8, 0.1)$ for the wave operator $\mathcal{L}u = u_{tt} - 4u_{xx}$, overlaid with partition blocks (left), and the pointwise error at the slice compared with the exact Green's function (right).

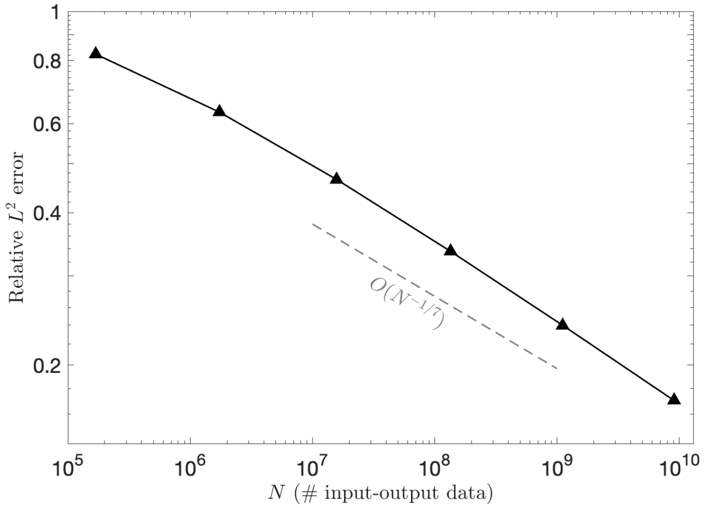

Figure 6: Empirical rate of convergence of Algorithm 2.

We do so via three theoretical contributions: (a) we prove new error bounds for the rSVD for HS operators in the operator norm, such that the suboptimality factor A_k of (17) tends to a constant as the target rank k tends to ∞; (b) we develop a scheme to detect the numerical rank of an operator using input-output products, with high probability; and (c) we introduce an adaptive component to the hierarchical partition of the Green's function's domain to effectively capture the location of its discontinuities.

In the remainder of this section, we discuss possible relaxations of our assumptions, a potential improvement to our scheme via the peeling algorithm, the difficulties of extending to higher-dimensional hyperbolic PDOs, and some challenges arising from the availability of input-output data.

6.1 Relaxations of assumptions

The main assumptions required of our unknown PDO $\mathcal{L}$ from (1a) are (a) homogeneous initial conditions, (b) self-adjointness, and (c) sufficient regularity of coefficients. Strict hyperbolicity is also required, but we do not discuss it because it enforces realistic physics by forbidding waves from traveling infinitely fast or backward in time.

Inhomogeneous initial conditions

For use as a direct solver, the homogeneous Green's function can be applied to solve Cauchy problems with certain inhomogeneous initial conditions, as a consequence of Duhamel's principle (Courant and Hilbert, 1962, Ch. VI.15). For initial conditions $u(x,0) = 0$ and $u_t(x,0) = \psi(x)$, one simply needs to integrate ψ against the Green's function on the line $\{t = 0\}$. If one also wants to solve initial value problems where $u(x,0)$ is not identically zero, then the values and derivatives of the coefficients of $\mathcal{L}$ must be given on $\{t = 0\}$ (Courant and Hilbert, 1962, Ch. V.5). We demonstrate the practical implementation of Duhamel's principle in Figure 7 by replacing the given inhomogeneous initial condition with an appropriate approximation of the Dirac delta on a short time strip, e.g., $f_\delta(x,t) := \delta^{-1}\chi_{[0,\delta]}(t)u_t(x,0)$, and setting our approximate solution to be

$$\tilde{u}(x,t) = \int_{D_T} \tilde{G}(x,t;y,s) f_\delta(y,s)\,\mathrm{d}y\,\mathrm{d}s, \qquad (x,t) \in D_T.$$

The inaccuracies seen in Figure 7 are due primarily to the use of an approximate Green's function. Because of the large amount of training data needed here, we are unable to construct a Green's function with sufficiently high accuracy with our available software and hardware.

One may be tempted to extrapolate the learned solution operator beyond the given time horizon by taking the value of the computed solution at the end of the first time interval to be the initial condition for the next time interval. However, without additional information about the coefficients of $\mathcal{L}$, one has no information whatsoever about the behavior of the Green's function beyond the time horizon on which the input-output training data is given. If it is known that the wave speed, for instance, is independent of time, then such extrapolations beyond the initial domain may be feasible.

Non-self-adjoint PDOs

While self-adjointness is not strictly necessary, more training data may be needed to learn the solution operator of a non-self-adjoint PDO to the same accuracy. This is because the coefficients of the adjoint PDO may have less regularity than the original operator. For instance, the wave operator with variable damping

$$\mathcal{L}u = u_{tt} - u_{xx} + k(x,t)u_t$$

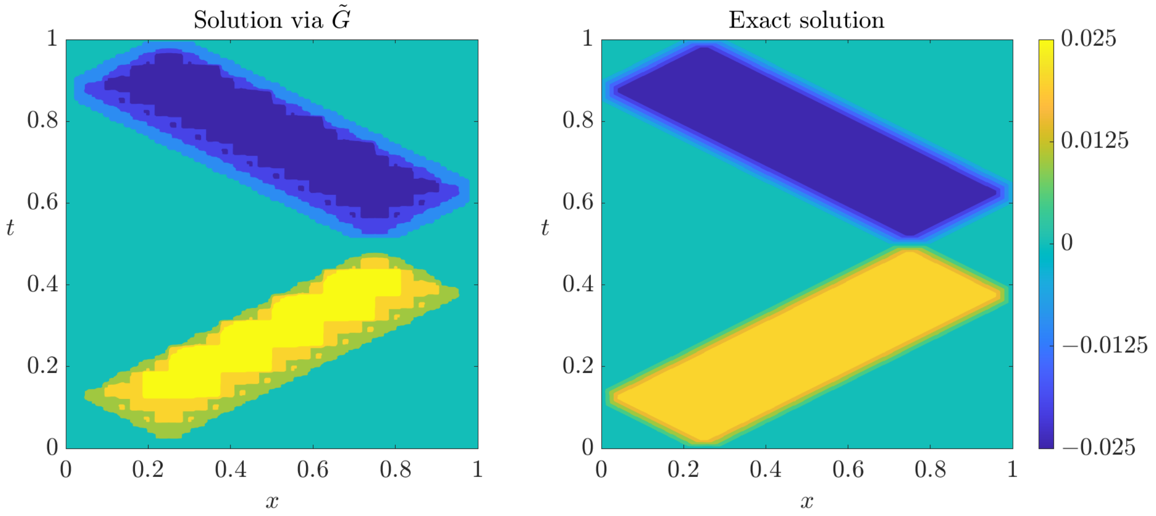

Figure 7: Approximate solution (left) and exact solution (right) to the wave operator $\mathcal{L}u = u_{tt} - 4u_{xx}$ with homogeneous boundary conditions but inhomogeneous, discontinuous initial conditions given by $u(x, 0) = 0$ and $u_t(x, 0) = \chi_{[0.2,0.3]}(x)$. The approximation was computed by multiplying the approximate Green's function $\tilde{G}$ derived from Algorithm 2 against an approximation of the Dirac delta on $\{t = 0\}$.

has adjoint

$$\mathcal{L}^* v = v_{tt} - v_{xx} - k(x,t)v_t - k_t(x,t)v.$$

If $k(x,t)$ is not smooth, then one of the coefficients of $\mathcal{L}^*$ has one less degree of regularity. Consequently, the corresponding homogeneous Green's function also has one less degree of regularity in the region of its support and thus requires more training data to recover (see Section A.5).

Whether or not our hyperbolic PDO is self-adjoint, in general one must have access to input-output data from both the Cauchy problem (1) as well as the corresponding adjoint problem (see footnote 6). Indeed, one cannot expect to recover an operator without the action of the adjoint (Boullé et al., 2024; Halikias and Townsend, 2024; Otto et al., 2023). In the special case where the hyperbolic PDO is self-adjoint and has time-independent coefficients, the homogeneous Cauchy problem is also self-adjoint.

Less regular coefficients

We assume throughout the paper that the coefficients of $\mathcal{L}$ are at least of class $\mathcal{C}^1$. It is certainly possible that our method extends to coefficients of lower regularity, but the Green's function theory for hyperbolic PDEs with low regularity coefficients is more subtle and may require one to work instead with parametrices. For example, Smith (1998) describes a parametrix for hyperbolic PDEs with Lipschitz coefficients whose spatial derivatives are also Lipschitz. For equations with even lower regularity, it is not always the case that the Cauchy problem is well-posed, and the regularity in the time and the space variables must be considered separately (see, e.g., Cicognani, 1999; Cicognani and Lorenz, 2018; Hurd and Sattinger, 1968; Reissig, 2003). It is unclear in such cases whether or not a Green's function approach, such as the one taken in this paper, is possible.

6.2 Peeling algorithm

Our algorithm requires performing rank detection and the rSVD on $O(\epsilon^{-6})$ subdomains. These operations require $O(\Psi_\epsilon^{-1}\epsilon^{-1}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ input-output training pairs for each subdomain; considering the subdomains one by one leads us to the number of training pairs described in Theorem 6.

However, it was recently shown in Boullé et al. (2023) that one can dramatically reduce the required number of input-output pairs by considering multiple subdomains simultaneously via the peeling algorithm (Lin et al., 2011). The argument relies in part on bounding the chromatic number of a graph associated with the hierarchical partition (Levitt and Martinsson, 2024). In our setting, the chromatic number at a given level L of the partition, denoted by $\chi(L)$, is approximately the number of subdomains of minimal volume—that is, cubes with side length 2^{-L}—seen in a 2-dimensional slice of the domain $D_T \times D_T$. For instance, the chromatic number associated with the partition at level $L = 3$ seen in Figure 4b is 52. Heuristically, the chromatic number is maximized when the characteristics are straight lines of some minimal slope $0 < m \leq 1$, in which case $\chi(L) \approx 2^L m^{-1}$. If the arguments in Boullé et al. (2023) can be repeated for hyperbolic PDOs, then the required number of input-output training pairs in Theorem 6 improves from $O(\Psi_\epsilon^{-1}\epsilon^{-7}\log(\Xi_\epsilon^{-1}\epsilon^{-1}))$ to $O(\Psi_\epsilon^{-1}\epsilon^{-3}\,\text{polylog}(\Xi_\epsilon^{-1}\epsilon^{-1}))$. Moreover, by Remark 7, this term continues to improve as one assumes greater regularity of the coefficients a, b, c of (1a). The challenge, which is addressed in Boullé et al. (2023) for elliptic PDEs, is that the peeling algorithm of Lin et al. (2011) is not proven to be stable. Hence, careful analysis is needed to show that errors from low-rank approximation and rank detection do not accumulate.

Notice that, unlike the elliptic case, the use of peeling cannot entirely reduce the ϵ^{-6} factor to $\text{polylog}(\epsilon^{-1})$. This is because the singularities of the Green's functions of hyperbolic PDEs in d variables occupy a $(2d-1)$-dimensional hypersurface in the $2d$-dimensional domain, due to the characteristics. In contrast, the singularities of the Green's functions of elliptic PDEs lie only on the diagonal, which is d-dimensional. In other words, the Green's functions of hyperbolic PDEs have an additional $d-1$ dimensions worth of singularities, and the chromatic number $\chi(L)$ thus grows exponentially with L. It is therefore doubtful that one can recover the Green's functions of hyperbolic PDEs at a sublinear rate without a significant development in the approximation theory of hierarchical matrices.

6.3 Extending to higher dimensions

We learn the Green's function of a hyperbolic PDE in two variables by detecting the location of its characteristics, which feature as jump discontinuities for the Green's function. A similar scheme should also work for detecting the singularities of Green's functions for hyperbolic PDEs in higher dimensions.

The main difficulty of higher-dimensional Green's functions is that they are not necessarily square-integrable. For instance, the Green's function of the wave equation in three spatial dimensions is given by

$$G(x,t;y,s) = \frac{\delta(t - s - \|x - y\|)}{4\pi\|x - y\|}, \qquad (x,t),(y,s) \in \mathbb{R}^3 \times [0,\infty),$$

where δ is the Dirac delta. Here, G has singularities of a distributional nature on the characteristics, which form a 7-dimensional hypersurface in the 8-dimensional domain of the Green's function. In this case, the theory of the rSVD for HS operators no longer applies since HS integral operators must have a square-integrable kernel. This difficulty has been circumvented to a degree in the case of parabolic PDEs, whose Green's functions are integrable but not square-integrable on the diagonal (Boullé et al., 2022b). Nevertheless, it is not immediately clear that the analysis carries over when the singularities are distributional.

Another factor that may play a role in higher dimensions is the effect of Huygens' principle (Courant and Hilbert, 1962; Evans, 2010). For the standard wave equation in d spatial dimensions, Huygens' principle implies that the support of the Green's function lies exclusively on the boundary of the future light cone when d is odd. In contrast, the support is the entire future light cone when d is even. For general hyperbolic PDOs, this effect does not correspond so nicely with the parity of the spatial dimension (see, e.g., Günther, 1991). Determining the impact of Huygens' principle may be an essential aspect of hyperbolic PDE learning in higher dimensions.

6.4 Availability of data

In practice, one often does not have access to exact solutions of the PDE in question. Instead, input-output training data may be generated from simulations driven by a numerical PDE solver. In such cases, the training data inherits whichever discretization errors the solver itself exhibits.

Figure 8 displays the output of our algorithm for training data coming from two different numerical solvers. The first is a first-order accurate upwind method (Banks and Henshaw, 2012) while the second is the simple second-order centered-in-time and centered-in-space finite difference method; both advance by explicit time-steps on a fixed-resolution uniform grid. Although much more advanced solvers exist, we chose these methods for their speed and simplicity given the large number of numerical solves required by our algorithm. Due to the numerical dissipation of the first-order method, we observe that the approximate Green's function is unable to locate the characteristics precisely, since they get smoothed out. On the other hand, the numerical dispersion exhibited by the second-order method fabricates artificial oscillations both inside and outside the light cones; these oscillations mislead our rank detection scheme into believing the Green's function is high-rank in regions where it is not, resulting in excessive partitioning within the light cones.

As such, the availability and accuracy of input-output data poses a major challenge for operator learning for hyperbolic PDEs that we encourage future research—both theoretical and practical—to address. While elliptic and parabolic PDEs significantly smooth out errors from the training data, those errors are structurally propagated throughout the domain in the hyperbolic case and must be dealt with in a careful manner. One possible approach is to use our algorithm to quickly construct a low-accuracy approximation of the Green's function to get a rough sense of where the characteristics lie. Then one can use this information to update the numerical solver generating the input-output data to better track propagation of singularities and minimize dispersion and dissipation errors. Alternating between using the Green's function approximation to improve the solver and vice versa may be an iterative solution to the problems presented above.

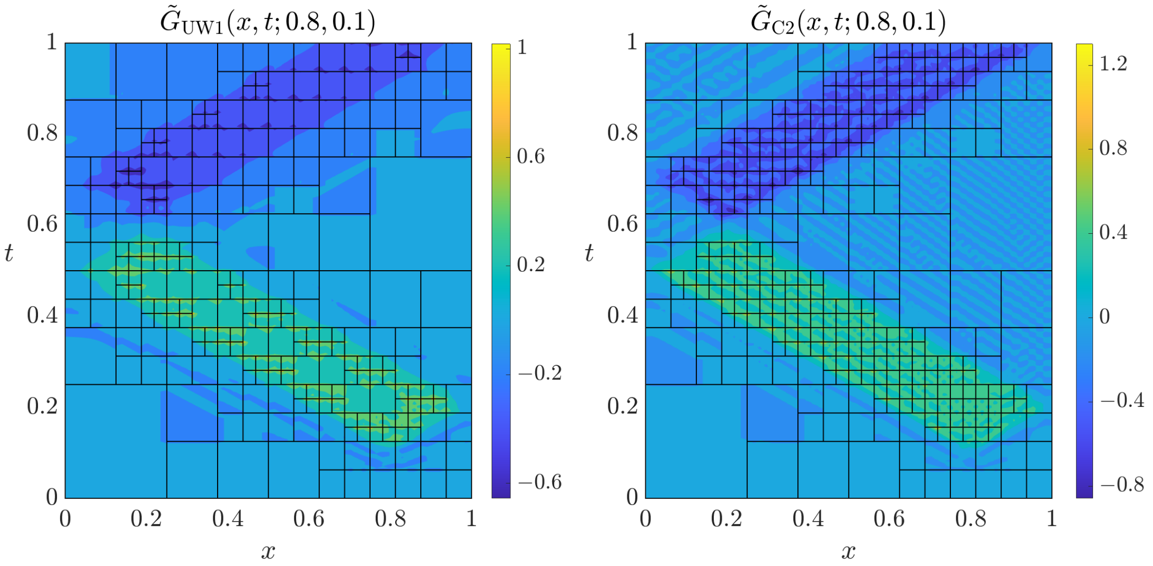

Figure 8: Approximate Green's function at the slice $(y,s)=(0.8,0.1)$ using input-output data generated by time-stepping numerical solvers: the first-order upwind method of Banks and Henshaw (2012) (left) and the second-order centered difference discretization (right).

Acknowledgments

The work of C. W. was supported by the NSF GRFP under grant DGE-2139899, and by the NSF under grant DMS-1929284 while the author was in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the Numerical PDEs: Analysis, Algorithms, and Data Challenges semester program. The work of A. T. was supported by the Office of Naval Research (ONR) under grant N00014-23-1-2729, the NSF under grant DMS-2045646, and a Weiss Junior Fellowship Award. Many thanks to Diana Halikias, Sungwoo Jeong, and Seick Kim for helpful discussions and suggestions.

Appendix A. Background Material

We review the relevant background material, comprising of HS operators, quasimatrices, orthogonal projectors, Gaussian processes, and approximation using Legendre series (Halko et al., 2011; Hsing and Eubank, 2015; Townsend and Trefethen, 2015; Trefethen, 2019; Vershynin, 2018).

A.1 Hilbert–Schmidt operators

Hilbert–Schmidt (HS) operators acting on L^2 functions are infinite-dimensional analogues of matrices acting on vectors. For $d \geq 1$, let $D_1, D_2 \subseteq \mathbb{R}^d$ be domains and let $L^2(D_i)$ be the space of square-integrable functions defined on D_i, for $i = 1, 2$, equipped with the inner product $\langle f, g\rangle := \int_{D_i} f(x)g(x)\,\mathrm{d}x$. Since $L^2(D_i)$ are separable Hilbert spaces, then for any compact linear operator $\mathscr{F} : L^2(D_1) \to L^2(D_2)$, there exists by the spectral theorem complete orthonormal bases $\{v_j\}_{j=1}^{\infty}$ and $\{e_j\}_{j=1}^{\infty}$ for $L^2(D_1)$ and $L^2(D_2)$ respectively, as well as a sequence of positive numbers $\sigma_1 \geq \sigma_2 \geq \cdots > 0$ such that

$$\mathscr{F} f = \sum_{\substack{j=1\\ \sigma_j>0}}^{\infty} \sigma_j \langle v_j, f\rangle e_j, \qquad f \in L^2(D_1), \tag{33}$$

where equality and convergence hold in the L^2 sense. This representation is called the singular value expansion (SVE) of $\mathscr{F}$; we call $\{e_j\}$ and $\{v_j\}$ the left and right singular functions of $\mathscr{F}$, respectively, and we call $\{\sigma_j\}$ the singular values of $\mathscr{F}$. The subspaces spanned by the first k left or right singular functions, for some positive integer k, are called dominant left and right k-singular subspaces of $\mathscr{F}$, respectively.

We say that $\mathscr{F}$ is a HS operator if its HS norm is finite, i.e.,

$$\|\mathscr{F}\|_{\mathrm{HS}} := \left(\sum_{j=1}^{\infty} \|\mathscr{F} v_j\|^2_{L^2(D_2)}\right)^{1/2} < \infty.$$

Since $\|\mathscr{F}\|^2_{\mathrm{HS}} = \sum_{j=1}^{\infty} \sigma_j^2$, the HS norm is an infinite-dimensional analogue of the Frobenius norm for matrices. We also have $\|\mathscr{F}\| \leq \|\mathscr{F}\|_{\mathrm{HS}}$. A special type of HS operator is a HS integral operator, which requires that $\mathscr{F}$ have an integral representation given by

$$(\mathscr{F} f)(x) = \int_{D_1} G(x, y) f(y)\,\mathrm{d}y, \qquad f \in L^2(D_1),\ x \in D_2. \tag{34}$$

for some kernel $G \in L^2(D_2 \times D_1)$. In this case, $\|\mathscr{F}\|_{\mathrm{HS}} = \|G\|_{L^2(D_2\times D_1)}$.

The adjoint of $\mathscr{F}$ is the unique compact linear operator $\mathscr{F}^* : L^2(D_2) \to L^2(D_1)$ satisfying $(\mathscr{F}^*)^* = \mathscr{F}$, $\|\mathscr{F}\| = \|\mathscr{F}^*\|$, and $\langle \mathscr{F} f, h\rangle_{L^2(D_2)} = \langle f, \mathscr{F}^* h\rangle_{L^2(D_1)}$ for every $f \in L^2(D_1)$ and $h \in L^2(D_2)$. The adjoint $\mathscr{F}^*$ has a SVE given by

$$\mathscr{F}^* g = \sum_{\substack{j=1\\ \sigma_j>0}}^{\infty} \sigma_j \langle e_j, g\rangle v_j, \qquad g \in L^2(D_2),$$

and for any $f \in L^2(D_1)$ and $h \in L^2(D_2)$ satisfies

$$(\mathscr{F}^*\mathscr{F}) f = \sum_{j=1}^{\infty} \sigma_j^2 \langle v_j, f\rangle v_j, \qquad (\mathscr{F}\mathscr{F}^*) h = \sum_{j=1}^{\infty} \sigma_j^2 \langle e_j, h\rangle e_j. \tag{35}$$

By the Eckart–Young–Mirsky theorem (Stewart and Sun, 1990, Thm. 4.18), truncating the SVE of a HS operator after k terms yields a best rank-k approximation in the HS and operator norms. If $\mathscr{F}_k : L^2(D_1) \to L^2(D_2)$ is defined by the truncation $\mathscr{F}_k f := \sum_{j=1}^{k} \sigma_j \langle v_j, f\rangle e_j$, then

$$\min_{\operatorname{rank}(\tilde{\mathscr{F}})=k} \|\mathscr{F} - \tilde{\mathscr{F}}\|_{\mathrm{HS}} = \|\mathscr{F} - \mathscr{F}_k\|_{\mathrm{HS}} = \left(\sum_{j=k+1}^{\infty} \sigma_j^2\right)^{1/2} \tag{36}$$

as well as

$$\min_{\operatorname{rank}(\tilde{\mathscr{F}})=k} \|\mathscr{F} - \tilde{\mathscr{F}}\| = \|\mathscr{F} - \mathscr{F}_k\| = \sigma_{k+1}. \tag{37}$$

Finally, we define the numerical rank of $\mathscr{F}$ with error tolerance $\epsilon > 0$, denoted by $\operatorname{rank}_\epsilon(\mathscr{F})$, to be the smallest integer k such that $\|\mathscr{F} - \mathscr{F}_k\| < \epsilon\|\mathscr{F}\|$. In particular, (37) implies that $\operatorname{rank}_\epsilon(\mathscr{F}) \leq k$ holds if $\sigma_{k+1} \leq \epsilon\sigma_1$.

A.2 Quasimatrices

Quasimatrices are infinite-dimensional analogues of tall-skinny matrices, and we use them principally to simplify notation and emphasize the analogy with matrices. A $D_1 \times k$ quasimatrix $\mathbf{\Omega}$ is represented by k columns, where each column is a function in $L^2(D_1)$, via

$$\mathbf{\Omega} = \begin{bmatrix}\omega_1 & \cdots & \omega_k\end{bmatrix}, \qquad \omega_j \in L^2(D_1).$$

One can similarly define $\infty \times k$ infinite matrices, in which each of the k columns is a sequence in ℓ^2, and likewise $D_1 \times \infty$ quasimatrices.

Many operations for rectangular matrices generalize to quasimatrices. If $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ is a HS operator, then $\mathscr{F}\mathbf{\Omega}$ denotes the quasimatrix given by applying $\mathscr{F}$ to each column of $\mathbf{\Omega}$, i.e.,

$$\mathscr{F}\mathbf{\Omega} = \begin{bmatrix}\mathscr{F}\omega_1 & \cdots & \mathscr{F}\omega_k\end{bmatrix}.$$

Quasimatrices have a transpose, denoted $\mathbf{\Omega}^*$, in the sense that

$$\mathbf{\Omega}^*\mathbf{\Omega} = \begin{bmatrix}\langle\omega_1,\omega_1\rangle & \cdots & \langle\omega_1,\omega_k\rangle \\ \vdots & \ddots & \vdots \\ \langle\omega_k,\omega_1\rangle & \cdots & \langle\omega_k,\omega_k\rangle\end{bmatrix}, \qquad \mathbf{\Omega}\mathbf{\Omega}^*(x,y) = \sum_{j=1}^{k} \omega_j(x)\omega_j(y).$$

Here, $\mathbf{\Omega}^*\mathbf{\Omega}$ is a $k \times k$ matrix of real numbers, while $\mathbf{\Omega}\mathbf{\Omega}^*$ is a function in $L^2(D_1 \times D_1)$ and can be thought of as an integral operator by taking $\mathbf{\Omega}\mathbf{\Omega}^*(x,y)$ to be its kernel. Quasimatrices can be thought of as HS operators on the relevant Hilbert spaces if, for instance, their singular values are square summable.

A.3 Orthogonal projectors

In analogy with the finite-dimensional setting, an orthogonal projection is a bounded self-adjoint linear operator $\mathbf{P} : L^2(D_2) \to L^2(D_2)$ that satisfies $\mathbf{P}^2 = \mathbf{P}$. Orthogonal projectors are completely determined by their range: for any closed subspace $\mathbf{M} \subseteq L^2(D_2)$, there is a unique orthogonal projector $\mathbf{P}_\mathbf{M}$ whose range is $\mathbf{M}$. This operator is compact if and only if $\mathbf{M}$ is finite-dimensional. For any $D_2 \times k$ quasimatrix $\mathbf{\Omega}$, the unique orthogonal projector associated with the column space of $\mathbf{\Omega}$ is denoted by

$$\mathbf{P}_{\mathbf{\Omega}} := \mathbf{\Omega}(\mathbf{\Omega}^*\mathbf{\Omega})^\dagger\mathbf{\Omega}^* : L^2(D_2) \to L^2(D_2),$$

as $\mathbf{P}_{\mathbf{\Omega}}\mathscr{F} : L^2(D_1) \to L^2(D_2)$ captures the orthogonal projection of the range of $\mathscr{F}$ onto the finite-dimensional column space of $\mathbf{\Omega}$. If $\mathbf{\Omega}$ has full column rank, then $\mathbf{\Omega}^*\mathbf{\Omega}$ is an invertible $k \times k$ matrix, and $(\mathbf{\Omega}^*\mathbf{\Omega})^\dagger = (\mathbf{\Omega}^*\mathbf{\Omega})^{-1}$.

We rely on the following inequality, analogous to that of Halko et al. (2011, Prop. 8.6), for the operator norm of orthogonal projectors.

Proposition 8 *Let $D_1, D_2 \subseteq \mathbb{R}^d$ be domains. Let $\mathbf{P} : L^2(D_2) \to L^2(D_2)$ be an orthogonal projector, and let $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ be a HS operator. Then*

$$\|\mathbf{P}\mathscr{F}\| \leq \|\mathbf{P}(\mathscr{F}\mathscr{F}^*)^q\mathscr{F}\|^{1/(2q+1)} \tag{38}$$

holds for every $q > 0$.

Proof The proof closely follows that of Halko et al. (2011, Prop. 8.6), making use of the Courant–Fischer minimax principle (Hsing and Eubank, 2015, Thm. 4.2.7). ■

A.4 Gaussian processes

A Gaussian process (GP) is an infinite-dimensional analogue of a multivariate Gaussian distribution, and a function drawn from a GP is analogous to a random vector drawn from a Gaussian distribution. For a domain $D \subseteq \mathbb{R}^d$ and a continuous symmetric positive semidefinite kernel $K : D \times D \to \mathbb{R}$, we define a GP with mean $\mu : D \to \mathbb{R}$ and covariance K to be a stochastic process $\{X_t : t \in D\}$ such that the random vector $(X_{t_1}, \ldots, X_{t_n})$, for every finite set of indices $t_1, \ldots, t_n \in D$, is a multivariate Gaussian distribution with mean $(\mu(t_1), \ldots, \mu(t_n))$ and covariance matrix $K_{ij} = K(t_i, t_j)$, $1 \leq i, j \leq n$. We denote such a GP by $\mathcal{GP}(\mu, K)$. Functions drawn from a GP with continuous kernel are almost surely in L^2. Additionally, since K is positive semidefinite, it has non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ with corresponding eigenfunctions $\{r_j\}_{j=1}^\infty$ that form an orthonormal basis for $L^2(D)$ (Hsing and Eubank, 2015, Thm. 4.6.5). The Karhunen–Lòeve expansion provides a method for sampling functions from a GP, as $f \sim \mathcal{GP}(0, K)$ has the expansion $f(t) = \sum_{j=1}^\infty \sqrt{\lambda_j} c_j r_j(t)$ converging in L^2 uniformly in $t \in D$, where $\{c_j\}_{j=1}^\infty$ are independent standard Gaussian random variables (Hsing and Eubank, 2015, Thm. 7.3.5). Additionally, we define the trace of K by $\mathrm{Tr}(K) := \sum_{j=1}^\infty \lambda_j < \infty$.

A.5 Low-rank approximation in the L^2 norm

In recovering a low-rank approximation of a kernel of a HS integral operator, we need explicit estimates on the decay of the kernel's singular values. We can obtain them by analyzing the kernel expressed in a Legendre series. Let $P_n^*(x)$ denote the shifted Legendre polynomial of degree n in the domain $[0, 1]$.[11] The shifted Legendre polynomials form an orthogonal basis

[11] Legendre polynomials are defined in NIST DLMF, Table 18.3.1. The shifted Legendre polynomials are simply the composition of Legendre polynomials with the transformation $x \mapsto 2(x-1)$ (see NIST DLMF, Ch. 18).

of $L^2([0,1])$, and if a function $f \in L^2([0,1])$ is uniformly Hölder continuous with parameter $> 1/2$, then it has a uniformly convergent Legendre expansion $f(x) = \sum_{n=0}^{\infty} a_n P_n^*(x)$. The truncated Legendre series $f_k(x) := \sum_{n=0}^{k} a_n P_n^*(x)$, for some $k \geq 0$, is the degree-k L^2 projection of f onto the space of polynomials of degree $\leq k$.

If f is r-times differentiable and $f^{(r)}$ has bounded total variation V, then for any $k > r+1$, we have

$$\|f - f_{k-1}\|_{L^2} \leq \frac{\sqrt{2}\,V}{\sqrt{\pi(r+1/2)}(k-r-1)^{r+1/2}} = O(k^{-(r+1/2)}), \tag{39}$$

where f_{k-1} is the degree-$(k-1)$ L^2 projection of f (Wang, 2023, Thm. 3.5). Similar results hold for a multivariate function $f(x_1, \ldots, x_n)$ if the conditions above hold uniformly over every 1-dimensional slice of f, which holds if f is r-times differentiable and $D^r f$ has total variation bounded by $V < \infty$ uniformly over all order-r partial derivatives D^r (Shi and Townsend, 2021; Townsend, 2014). In particular, for any positive integers m, n, we may consider $f : [0,1]^{m+n} \to \mathbb{R}$ and its degree-$(k-1)$ L^2 projection f_{k-1} as the kernels of HS integral operators $\mathscr{F}, \tilde{\mathscr{F}} : L^2([0,1]^m) \to L^2([0,1]^n)$, respectively. Since f_{k-1} is a polynomial of degree $k-1$, then $\tilde{\mathscr{F}}$ is a rank-k operator. By the Eckart–Young–Mirsky theorem,

$$\left(\sum_{j=k+1}^{\infty} \sigma_j^2\right)^{1/2} \leq \|\mathscr{F} - \tilde{\mathscr{F}}\|_{\mathrm{HS}} = \|f - f_{k-1}\|_{L^2} \leq \frac{\sqrt{2}\,V}{\sqrt{\pi(r+1/2)}(k-r-1)^{r+1/2}},$$

where σ_j denotes the jth singular value of $\mathscr{F}$. Combining this with the observation that $k\sigma_{2k}^2 \leq \sum_{j=k+1}^{2k} \sigma_j^2 \leq \sum_{j=k+1}^{\infty} \sigma_j^2$, we obtain an explicit decay rate for the singular values of $\mathscr{F}$, i.e.,

$$\sigma_k \leq \frac{2^{r+3/2} V}{\sqrt{\pi(r+1/2)} \cdot \sqrt{k}(k-2r-2)^{r+1/2}} = O(k^{-(r+1)}) \tag{40}$$

for $k > 2(r+1)$.

Appendix B. Randomized Singular Value Decomposition

This appendix contains proofs of results relating to the rSVD in Section 3.

B.1 Proofs for Section 3.1

In this section, we prove Theorems 1 and 3, which bound the error of an approximant generated by the rSVD with respect to the operator norm. The proof is similar to that of Halko et al. (2011, Thms. 10.6 and 10.8), generalized to the infinite-dimensional setting with non-standard Gaussian vectors.

We first introduce notation. For separable Hilbert spaces $\mathcal{H}, \mathcal{K}$, let $\mathrm{HS}(\mathcal{H}, \mathcal{K})$ be the Hilbert space of HS operators $\mathcal{H} \to \mathcal{K}$. Given a bounded, self-adjoint linear operator $\mathbf{C} : \mathcal{K} \to \mathcal{K}$, we define a quadratic form on $\mathcal{K}$ by

$$\langle x, y \rangle_{\mathbf{C}} := \langle x, \mathbf{C}y \rangle_{\mathcal{K}}, \qquad x, y \in \mathcal{K},$$

as well as a quadratic form on $\mathrm{HS}(\mathcal{H},\mathcal{K})$ by

$$\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{C}} := \langle \mathbf{A}, \mathbf{C}\mathbf{B}\rangle_{\mathrm{HS}}.$$

If $\mathbf{C}$ is positive definite, then both quadratic forms are inner products on $\mathcal{K}$ and $\mathrm{HS}(\mathcal{H},\mathcal{K})$, respectively. Then we let $\|x\|_{\mathbf{C}}^2 := \langle x, x\rangle_{\mathbf{C}}$ and $\|\mathbf{A}\|_{\mathbf{C}}^2 := \langle \mathbf{A}, \mathbf{A}\rangle_{\mathbf{C}}$ denote the induced norms. We remark that $\|x\|_{\mathbf{C}} = \|\mathbf{C}^{1/2}x\|_{\mathcal{K}}$ and $\|\mathbf{A}\|_{\mathbf{I}\to\mathbf{C}} = \|\mathbf{C}^{1/2}\mathbf{A}\|$, where $\|\mathbf{A}\|_{\mathbf{I}\to\mathbf{C}}$ denotes the operator norm of $\mathbf{A}$ when viewed as an operator $(\mathcal{H}, \|\cdot\|_{\mathcal{H}}) \to (\mathcal{K}, \|\cdot\|_{\mathbf{C}})$. Finally, we let $\mathcal{H}\otimes\mathcal{K}$ denote the tensor product of Hilbert spaces, which embeds into $\mathrm{HS}(\mathcal{H},\mathcal{K})$ by associating with each $x\otimes y \in \mathcal{H}\otimes\mathcal{K}$ the operator $u \mapsto \langle x, u\rangle_{\mathcal{H}}\, y$.

Lemma 9 *Let $\mathcal{H},\mathcal{K}$ be Hilbert spaces. For any $u, w \in \mathcal{H}$ and $v, z \in \mathcal{K}$, we have*

$$\|u\otimes v - w\otimes z\|_{\mathrm{HS}}^2 \le \max(\|u\|^2, \|w\|^2)\|v-z\|^2 + \max(\|v\|^2, \|z\|^2)\|u-w\|^2.$$

Proof Observe that $\langle u\otimes v, w\otimes z\rangle_{\mathrm{HS}} = \langle u, w\rangle\langle v, z\rangle$. Then

$$\begin{aligned}\|u\otimes v - w\otimes z\|^2 &= \|u\otimes v\|^2 + \|w\otimes z\|^2 - 2\langle u\otimes v, w\otimes z\rangle\\ &= \|u\|^2\|v\|^2 + \|w\|^2\|z\|^2 - 2\langle u, w\rangle\langle v, z\rangle.\end{aligned}$$

Let $A = \max(\|v\|, \|z\|)$ and $B = \max(\|u\|, \|w\|)$. Then

$$\begin{aligned}&A^2\|u-w\|^2 + B^2\|v-z\|^2\\ &\qquad = A^2(\|u\|^2 + \|w\|^2) + B^2(\|v\|^2 + \|z\|^2) - 2(A^2\langle u, w\rangle + B^2\langle v, z\rangle).\end{aligned}$$

Observe that $(B^2 - \langle u, w\rangle)(A^2 - \langle v, z\rangle) \ge 0$, hence

$$\|u\otimes v - w\otimes z\|_{\mathrm{HS}}^2 \le \|u\|^2\|v\|^2 + \|w\|^2\|z\|^2 + 2A^2B^2 - 2(A^2\langle u, w\rangle + B^2\langle v, z\rangle).$$

It is straightforward to verify that

$$\|u\|^2\|v\|^2 + \|w\|^2\|z\|^2 + 2A^2B^2 \le A^2(\|u\|^2 + \|w\|^2) + B^2(\|v\|^2 + \|z\|^2),$$

which completes the proof. ■

We extend a result of Halko et al. (2011, Prop. 10.1) to the infinite-dimensional setting, following Gordon (1985, 1988). To simplify notation, we use $\|\cdot\|$ to denote either the operator, ℓ^2, or $L^2(D)$ norm depending on context. We also use $\mathcal{N}(0, \mathbf{I}\otimes\mathbf{C})$ to denote the distribution of a mean zero random matrix with independent columns drawn from $\mathcal{N}(0, \mathbf{C})$.

Proposition 10 *Let $\mathbf{G}$ be an $\infty\times n$ infinite random matrix with distribution $\mathcal{N}(0, \mathbf{I}_n\otimes\mathbf{C})$, where $\mathbf{C}$ is a symmetric positive definite $\infty\times\infty$ covariance matrix with $\mathrm{Tr}(\mathbf{C}) < \infty$. Let $\mathbf{S}$ be a $D\times\infty$ quasimatrix and let $\mathbf{T}$ be a $n\times\infty$ infinite matrix, both deterministic with $\|\mathbf{S}\|_{\mathrm{HS}}, \|\mathbf{T}\|_{\mathrm{HS}} < \infty$. Then*

$$\mathbb{E}\|\mathbf{S}\mathbf{G}\mathbf{T}\| \le \Big(\|\mathbf{C}^{1/2}\|_{\mathrm{HS}}\|\mathbf{T}\| + \|\mathbf{C}^{1/2}\|\|\mathbf{T}\|_{\mathrm{HS}}\Big)\|\mathbf{S}\|.$$

Proof Let $g \sim \mathcal{N}(0, \|\mathbf{T}\|^2\mathbf{C})$ be Gaussian in ℓ^2 and let $h \sim \mathcal{N}(0, \|\mathbf{C}^{1/2}\mathbf{S}\|^2\,\mathbf{I}_n)$ be a Gaussian random vector in $\mathbb{R}^n$, such that g, h, and $\mathbf{G}$ are all independent. Define

$$X(u,v) := \langle \mathbf{SGT}u, v\rangle, \quad Y(u,v) := \langle \mathbf{S}g, v\rangle + \langle \mathbf{T}^*h, u\rangle$$

indexed by $u \in \ell^2$ and $v \in L^2(D)$ of unit length. One can show that $X(u,v)$ and $Y(u,v)$ are well-defined, and that $X(\cdot,\cdot)$ and $Y(\cdot,\cdot)$ are mean zero Gaussian processes.

We aim to apply the Sudakov–Fernique inequality (Vershynin, 2018, Thm. 7.2.11). For any unit-length $u, w \in \ell^2$ and $v, z \in L^2(D)$, set $\mathbf{A} = (\mathbf{T}u) \otimes (\mathbf{S}^*v) - (\mathbf{T}w) \otimes (\mathbf{S}^*z)$. We compute

$$\begin{aligned}\mathbb{E}(X(u,v) - X(w,z))^2 &= \mathrm{Tr}(\mathbf{A}^*\mathbf{C}\mathbf{A})\\ &\le \|\mathbf{T}\|^2\|\mathbf{S}^*(v-z)\|_{\mathbf{C}}^2 + \|\mathbf{S}\|_{\mathbf{I}\to\mathbf{C}}^2\|\mathbf{T}(u-w)\|^2\\ &= \|\mathbf{T}\|^2\|\mathbf{C}^{1/2}\mathbf{S}^*(v-z)\|^2 + \|\mathbf{C}^{1/2}\mathbf{S}\|^2\|\mathbf{T}(u-w)\|^2. \qquad (41)\end{aligned}$$

The first equality follows from fully expanding $(X(u,v) - X(w,z))^2$ and using the definition of $\mathbf{G}$; the inequality is obtained by applying Lemma 9 taking $\mathcal{H}$ to be $\mathbb{R}^n$ with the usual inner product, and $\mathcal{K}$ to be ℓ^2 with the inner product induced by $\mathbf{C}$.

On the other hand, observe that

$$\mathbb{E}\langle \mathbf{S}g, v - z\rangle^2 = \|\mathbf{T}\|^2(v-z)^*\mathbf{SCS}^*(v-z) = \|\mathbf{T}\|^2\|\mathbf{C}^{1/2}\mathbf{S}^*(v-z)\|^2,$$

and similarly $\mathbb{E}\langle \mathbf{T}^*h, u - w\rangle^2 = \|\mathbf{C}^{1/2}\mathbf{S}\|^2\|\mathbf{T}(u-w)\|^2$. Therefore, by independence of g and h, we have

$$\mathbb{E}(Y(u,v) - Y(w,z))^2 = \|\mathbf{T}\|^2\|\mathbf{C}^{1/2}\mathbf{S}^*(v-z)\|^2 + \|\mathbf{C}^{1/2}\mathbf{S}\|^2\|\mathbf{T}(u-w)\|^2. \qquad (42)$$

Combining (41) and (42) yields $\mathbb{E}(X(u,v) - X(w,z))^2 \le \mathbb{E}(Y(u,v) - Y(w,z))^2$ for every unit-length u, v, w, z, so by the Sudakov–Fernique inequality, we have

$$\mathbb{E} \sup_{\substack{u\in\ell^2\\ \|u\|=1}} \sup_{\substack{v\in L^2(D)\\ \|v\|=1}} \langle \mathbf{SGT}u, v\rangle \le \mathbb{E} \sup_{\substack{u\in\ell^2\\ \|u\|=1}} \sup_{\substack{v\in L^2(D)\\ \|v\|=1}} (\langle \mathbf{S}g, v\rangle + \langle \mathbf{T}^*h, u\rangle). \qquad (43)$$

Since $\mathbf{SGT}$ is almost surely a bounded operator, then the left-hand side of (43) is $\mathbb{E}\|\mathbf{SGT}\|$. For the right-hand side, two applications of the Cauchy–Schwarz inequality yield

$$\mathbb{E} \sup_{\substack{v\in L^2(D)\\ \|v\|=1}} \langle \mathbf{S}g, v\rangle \le (\mathbb{E}\|\mathbf{S}g\|^2)^{1/2} = \mathrm{Tr}(\|\mathbf{T}\|^2\mathbf{SCS}^*)^{1/2} \le \|\mathbf{C}^{1/2}\|_{\mathrm{HS}}\|\mathbf{S}\|\|\mathbf{T}\|$$

An analogous argument for $\langle \mathbf{T}^*h, u\rangle$ combined with (43) completes the proof. ■

Finally, we derive probability estimates for the operator norm of the pseudoinverse of non-standard Gaussian matrices, following Chen and Dongarra (2005) and Edelman (1988).

Proposition 11 *Let $\mathbf{G}$ be an $m \times n$ random matrix with distribution $\mathcal{N}(0, \mathbf{I}_n \otimes \mathbf{C})$, where $\mathbf{C}$ is a symmetric positive definite $m \times m$ covariance matrix with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_m > 0$. If $n \ge m \ge 2$, then for any $t \ge 0$, we have*

$$\mathbb{P}\{\|\mathbf{G}^\dagger\| \ge t\} \le \frac{1}{\sqrt{2\pi(n-m+1)}}\left(\frac{e\sqrt{n}}{\sqrt{\lambda_m}(n-m+1)}\right)^{n-m+1} t^{-(n-m+1)}.$$

Proof Let $x_1 \geq \cdots \geq x_m > 0$ be the eigenvalues of the Wishart matrix $\mathbf{G}\mathbf{G}^* \sim \mathcal{W}_m(n, \mathbf{C})$. Since $\|\mathbf{G}^\dagger\| = x_m^{-1/2}$, then we focus on the distribution of x_m. It is well-known (James, 1964, 1968) that the joint density of the eigenvalues is given by

$$f(x_1, \ldots, x_m) = \frac{|\mathbf{C}|^{-\frac{n}{2}}}{K_{m,n}} {}_0F_0(\mathbf{C}^{-1}, -\tfrac{1}{2}\mathbf{X}) \prod_{1 \leq i < j \leq m} (x_i - x_j) \prod_{i=1}^{m} x_i^{\frac{n-m-1}{2}},$$

where $\mathbf{X} = \mathrm{diag}(x_1, \ldots, x_m)$, $|\cdot|$ is the determinant function, ${}_0F_0$ is the hypergeometric function of two matrix arguments, Γ_m is the multivariate Gamma function, and

$$K_{m,n} = \frac{\pi^{\frac{m^2}{2}}}{2^{\frac{mn}{2}} \Gamma_m(\frac{n}{2}) \Gamma_m(\frac{m}{2})}$$

(for details and definitions, see Muirhead, 1982). We bound the density function since the hypergeometric function is increasing in the eigenvalues of its arguments (Kates, 1981, Thm. IV.1), that is,

$${}_0F_0(\mathbf{C}^{-1}, -\tfrac{1}{2}\mathbf{X}) \leq {}_0F_0(\tfrac{1}{\lambda_m}\mathbf{I}_m, -\tfrac{1}{2}\mathbf{X}) = e^{-\frac{1}{2\lambda_m}\sum_{i=1}^{m} x_i}$$

where the equality follows from Muirhead (1982, Prob. 7.7). Therefore,

$$f(x_1, \ldots, x_m) \leq \frac{|\mathbf{C}|^{-\frac{n}{2}}}{K_{m,n}} e^{-\frac{1}{2\lambda_m}\sum_{i=1}^{m} x_i} \prod_{1 \leq i < j \leq m} (x_i - x_j) \prod_{i=1}^{m} x_i^{\frac{n-m-1}{2}}.$$

We bound the density function f_{x_m} of x_m by integration. Let $R_x = \{(x_1, \ldots, x_{m-1}) : x_1 \geq \cdots \geq x_{m-1} \geq x\}$, so

$$f_{x_m}(x) \leq \frac{|\mathbf{C}|^{-\frac{n}{2}}}{K_{m,n}} x^{\frac{n-m-1}{2}} e^{-\frac{x}{2\lambda_m}} \int_{R_x} e^{-\frac{1}{2\lambda_m}\sum_{i=1}^{m-1} x_i} \prod_{1 \leq i < j \leq m-1} (x_i - x_j) \prod_{i=1}^{m-1} (x_i - x) x_i^{\frac{n-m-1}{2}} \, \mathrm{d}x_i.$$

We bound further by using $x_i - x \leq x_i$ and the non-negativity of the integrand, then perform the change of variables $x_i = \lambda_m y_i$ to obtain

$$f_{x_m}(x) \leq \frac{|\mathbf{C}|^{-\frac{n}{2}}}{K_{m,n}} \lambda_m^{\frac{(m-1)(n+1)}{2}} x^{\frac{n-m-1}{2}} e^{-\frac{x}{2\lambda_m}} \int_{R_0} e^{-\frac{1}{2}\sum_{i=1}^{m-1} y_i} \prod_{1 \leq i < j \leq m-1} (y_i - y_j) \prod_{i=1}^{m-1} y_i^{\frac{n-m+1}{2}} \, \mathrm{d}y_i,$$

where $R_0 = \{(y_1, \ldots, y_{m-1}) : y_1 \geq \cdots \geq y_{m-1} \geq 0\}$. Notice that the integrand is simply the unnormalized joint density of the eigenvalues of a $\mathcal{W}_{m-1}(n+1, \mathbf{I}_{m-1})$ matrix, so the integral evaluates to $K_{m-1,n+1}$. By Chen and Dongarra (2005, Eq. (3.14)), we have

$$\frac{K_{m-1,n+1}}{K_{m,n}} = \frac{2^{\frac{n-m-1}{2}} \Gamma(\frac{n+1}{2})}{\Gamma(\frac{m}{2}) \Gamma(n-m+1)},$$

as well as

$$|\mathbf{C}|^{-\frac{n}{2}} \lambda_m^{\frac{(m-1)(n+1)}{2}} \leq \lambda_m^{\frac{m-n-1}{2}}.$$

Thus,

$$f_{x_m}(x) \leq \frac{2^{\frac{n-m-1}{2}} \lambda_m^{\frac{m-n-1}{2}} \Gamma(\frac{n+1}{2})}{\Gamma(\frac{m}{2})\Gamma(n-m+1)} x^{\frac{n-m-1}{2}} e^{-\frac{x}{2\lambda_m}},$$

and we obtain

$$\mathbb{P}\{x_m \leq x^{-2}\} \leq \frac{1}{\Gamma(n-m+2)} \left(\frac{\sqrt{n}}{x\sqrt{\lambda_m}}\right)^{n-m+1}$$

by following Chen and Dongarra (2005, Lem. 4.1). We conclude by applying Stirling's approximation. ■

Corollary 12 *Under the same assumptions as Proposition 11, we have*

$$\mathbb{E}\|\mathbf{G}^\dagger\| < \frac{e\sqrt{n}}{\sqrt{\lambda_m}(n-m)}.$$

Proof The proof is the same as that in Halko et al. (2011, Prop. A.4). ■

We are ready to prove Theorem 1.

Proof (Theorem 1) Given the notation of Section 3.1, let $\mathbf{\Omega}_1 = \mathbf{V}_1^*\mathbf{\Omega}$ and $\mathbf{\Omega}_2 = \mathbf{V}_2^*\mathbf{\Omega}$. By (Boullé and Townsend, 2023, Lem. 1)—which holds for any unitarily invariant norm—we have

$$\mathbb{E}\|\mathscr{F} - \mathbf{P}_\mathbf{Y}\mathscr{F}\| \leq \|\mathbf{\Sigma}_2\| + \mathbb{E}\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|.$$

To compute $\mathbb{E}\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|$, notice that $\mathbf{V}^*\mathbf{\Omega} \sim \mathcal{N}(0, \mathbf{I}_{k+p} \otimes \mathbf{C})$, by (Boullé and Townsend, 2023, Lem. 1), with columns that are almost surely in ℓ^2 since $\mathrm{Tr}(\mathbf{C}) = \mathrm{Tr}(K) < \infty$. Let $\mathbf{\Omega}_2|\mathbf{\Omega}_1$ denote the random matrix $\mathbf{\Omega}_2$ conditioned on $\mathbf{\Omega}_1$, which, by Mandelbaum (1984, Thm. 2 and Cor. 2), has distribution $\mathcal{N}(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{\Omega}_1, \mathbf{I}_{k+p} \otimes (\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}))$. Let $\bar{\mathbf{\Omega}}_2$ be normally distributed with mean zero and the same covariance as $\mathbf{\Omega}_2|\mathbf{\Omega}_1$, sampled independently of $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, so that $\mathbf{\Omega}_2|\mathbf{\Omega}_1 \sim \bar{\mathbf{\Omega}}_2 + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{\Omega}_1$. We now condition on $\mathbf{\Omega}_1$ to obtain, by Proposition 10,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\| \\
&\quad= \mathbb{E}\|\mathbf{\Sigma}_2(\bar{\mathbf{\Omega}}_2 + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{\Omega}_1)\mathbf{\Omega}_1^\dagger\| \\
&\quad\leq \left(\|(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})^{1/2}\|_{\mathrm{HS}}\, \mathbb{E}\|\mathbf{\Omega}_1^\dagger\| + \|\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|^{1/2}\, \mathbb{E}\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}}\right)\|\mathbf{\Sigma}_2\| \\
&\qquad + \mathbb{E}\|\mathbf{\Sigma}_2\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{\Omega}_1\mathbf{\Omega}_1^\dagger\|.
\end{aligned}$$

First, observe that $\mathbf{\Omega}_1$ has full rank with probability 1, so $\mathbf{\Omega}_1\mathbf{\Omega}_1^\dagger = \mathbf{I}_k$. Thus,

$$\mathbb{E}\|\mathbf{\Sigma}_2\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{\Omega}_1\mathbf{\Omega}_1^\dagger\| \leq \lambda_1\|\mathbf{\Sigma}_2\|\|\mathbf{C}_{11}^{-1}\|,$$

using the bounds $\|\mathbf{C}_{21}\| \leq \|\mathbf{C}\| \leq \lambda_1$. Additionally, since $\mathbf{C} \succeq \mathbf{C}_{22} \succeq \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$, where $\succeq$ denotes the Löwner order, then we have

$$\|(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})^{1/2}\|_{\mathrm{HS}} \leq \|\mathbf{C}^{1/2}\|_{\mathrm{HS}} = \sqrt{\mathrm{Tr}(\mathbf{C})} = \sqrt{\mathrm{Tr}(K)}$$

and $\|\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|^{1/2} \leq \sqrt{\lambda_1}$. Finally, by Corollary 12 and Boullé and Townsend (2023, Eq. (10)), we have

$$\mathbb{E}\|\mathbf{\Omega}_1^\dagger\| \leq \frac{e\sqrt{\|\mathbf{C}_{11}^{-1}\|(k+p)}}{p}, \qquad \mathbb{E}\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}} = \sqrt{\frac{\mathrm{Tr}(\mathbf{C}_{11}^{-1})}{p+1}},$$

which altogether yields

$$\mathbb{E}\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\| \leq \left(\frac{e\sqrt{\mathrm{Tr}(K)\|\mathbf{C}_{11}^{-1}\|(k+p)}}{p} + \sqrt{\frac{\lambda_1\,\mathrm{Tr}(\mathbf{C}_{11}^{-1})}{p+1}} + \lambda_1\|\mathbf{C}_{11}^{-1}\|\right)\|\mathbf{\Sigma}_2\|.$$

The estimate (13) for the average error in the operator norm now follows.

For (14), we follow Halko et al. (2011, Thm. 10.8) and define, for each $t \geq 1$, the event

$$E_t := \left\{\|\mathbf{\Omega}_1^\dagger\| \leq \frac{e\sqrt{\|\mathbf{C}_{11}^{-1}\|(k+p)}}{p+1}\cdot t \quad \text{and} \quad \|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}} \leq \sqrt{\frac{\mathrm{Tr}(\mathbf{C}_{11}^{-1})}{p+1}}\cdot t\right\}$$

such that $\mathbb{P}(E_t^c) \leq 2t^{-p}$ by Proposition 11 and Boullé and Townsend (2023, Lem. 3). We also consider the random function $h(\mathbf{X}) := \|\mathbf{\Sigma}_2\mathbf{X}\mathbf{\Omega}_1^\dagger\|$ defined on the space of $\infty \times (k+p)$ quasimatrices. The corresponding Cameron–Martin space $\mathcal{M}$ with respect to the distribution $\mathcal{N}(0, \mathbf{I}_{k+p}\otimes\mathbf{C})$ is the space of $\infty\times(k+p)$ quasimatrices $\mathbf{Y}$ satisfying $\mathrm{Tr}(\mathbf{Y}^*\mathbf{C}^{-1}\mathbf{Y}) < \infty$, or equivalently,

$$\mathcal{M} = \{\mathbf{C}^{1/2}\mathbf{X} : \mathbf{X} \text{ is an } \infty \times (k+p) \text{ quasimatrix}\},$$

equipped with the inner product $\langle \mathbf{Y}, \mathbf{Z}\rangle_{\mathcal{M}} := \mathrm{Tr}(\mathbf{Y}^*\mathbf{C}^{-1}\mathbf{Z})$ (Bogachev, 1998, Ch. 2). Notice that $\|\mathbf{Y}\|_{\mathrm{HS}} \leq \sqrt{\lambda_1}\|\mathbf{Y}\|_{\mathcal{M}}$ for every $\mathbf{Y} \in \mathcal{M}$. Thus, the function h conditioned on $\mathbf{\Omega}_1$ is $\mathcal{M}$-Lipschitzian with constant $\sqrt{\lambda_1}\|\mathbf{\Sigma}_2\|\|\mathbf{\Omega}_1^\dagger\|$, since

$$|h(\mathbf{X}+\mathbf{Y}) - h(\mathbf{X})| \leq \|\mathbf{\Sigma}_2\mathbf{Y}\mathbf{\Omega}_1^\dagger\| \leq \sqrt{\lambda_1}\|\mathbf{\Sigma}_2\|\|\mathbf{\Omega}_1^\dagger\|\|\mathbf{Y}\|_{\mathcal{M}}$$

holds for every quasimatrix $\mathbf{X}$ and every $\mathbf{Y} \in \mathcal{M}$. We now apply the concentration inequality of Bogachev (1998, Thm. 4.5.7) to h conditioned on $\mathbf{\Omega}_1$, to obtain, for every $s \geq 0$,

$$\mathbb{P}\left\{\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\| > \left(\sqrt{\mathrm{Tr}(K)}\|\mathbf{\Omega}_1^\dagger\| + \sqrt{\lambda_1}\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}} + \lambda_1\|\mathbf{C}_{11}^{-1}\| + \sqrt{\lambda_1}\|\mathbf{\Omega}_1^\dagger\|\cdot s\right)\|\mathbf{\Sigma}_2\| \mid E_t\right\} \leq e^{-s^2/2},$$

where we make use of the fact

$$\mathbb{E}[h(\mathbf{\Omega}_2) \mid \mathbf{\Omega}_1] \leq \left(\sqrt{\mathrm{Tr}(K)}\|\mathbf{\Omega}_1^\dagger\| + \sqrt{\lambda_1}\|\mathbf{\Omega}_1^\dagger\|_{\mathrm{F}} + \lambda_1\|\mathbf{C}_{11}^{-1}\|\right)\|\mathbf{\Sigma}_2\|$$

by Proposition 10. By definition of E_t, it follows that

$$\mathbb{P}\left\{\|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\| > \left[\left(\sqrt{\mathrm{Tr}(K)} + s\sqrt{\lambda_1}\right)\frac{e\sqrt{\|\mathbf{C}_{11}^{-1}\|(k+p)}}{p+1}\cdot t + \right.\right.$$
$$\left.\left. + \sqrt{\frac{\lambda_1\,\mathrm{Tr}(\mathbf{C}_{11}^{-1})}{p+1}}\cdot t + \lambda_1\|\mathbf{C}_{11}^{-1}\|\right]\|\mathbf{\Sigma}_2\| \mid E_t\right\} \leq e^{-s^2/2}.$$

Since $\mathbb{P}(E_t^c) \leq 2t^{-p}$, then we obtain

$$\mathbb{P}\left\{ \|\mathbf{\Sigma}_2\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\| > \left[\left(\sqrt{\operatorname{Tr}(K)} + s\sqrt{\lambda_1}\right) \frac{e\sqrt{\|\mathbf{C}_{11}^{-1}\|(k+p)}}{p+1} \cdot t + \right.\right.$$

$$\left.\left. + \sqrt{\frac{\lambda_1 \operatorname{Tr}(\mathbf{C}_{11}^{-1})}{p+1}} \cdot t + \lambda_1\|\mathbf{C}_{11}^{-1}\| \right] \|\mathbf{\Sigma}_2\| \right\} \leq 2t^{-p} + e^{-s^2/2}.$$

We combine this estimate with Boullé and Townsend (2023, Lem. 1) to conclude. ■

The proof of Theorem 3 follows easily.

Proof (Theorem 3) We first claim that $\|\mathscr{F} - \mathbf{P_Z}\mathscr{F}\| \leq \|\mathscr{H} - \mathbf{P_Z}\mathscr{H}\|^{1/(2q+1)}$. Indeed, let $\mathbf{I}$ be the identity operator on $L^2(D_2)$ and observe that $\mathbf{I} - \mathbf{P_Z}$ is an orthogonal projector. Then Proposition 8 yields

$$\|(\mathbf{I} - \mathbf{P_Z})\mathscr{F}\| \leq \|(\mathbf{I} - \mathbf{P_Z})(\mathscr{F}\mathscr{F}^*)^q\mathscr{F}\|^{1/(2q+1)} = \|(\mathbf{I} - \mathbf{P_Z})\mathscr{H}\|^{1/(2q+1)}.$$

Now noting that the singular values of $\mathscr{H}$ are $\sigma_1^{2q+1} \geq \sigma_2^{2q+1} \geq \cdots$, the result immediately follows from Theorem 1. ■

B.2 Proofs for Section 3.2

Here, prove Theorem 5. In the following, let $\mathbf{\Theta}(\mathcal{X}, \mathcal{Y})$ denote the diagonal matrix of canonical angles between two subspaces $\mathcal{X}, \mathcal{Y}$ of $L^2(D_2)$ with equal finite dimension (Stewart and Sun, 1990, Def. I.5.3). Additionally, for any (quasi)matrix $\mathbf{X}$, we denote its column space by the calligraphic version of the same symbol, namely, $\mathcal{X}$.

Lemma 13 *Let $\mathscr{F} : L^2(D_1) \to L^2(D_2)$ be a HS operator with SVE given by (33). Select an integer $k \geq 1$ such that $\sigma_k > \sigma_{k+1}$. Select an integer $q \geq 0$ and set*

$$\delta_q = \left[\left(\frac{\sigma_k}{\sigma_{k+1}}\right)^{2q+1} - 1\right]^{-1} > 0. \tag{44}$$

Let $\mathscr{H} := (\mathscr{F}\mathscr{F}^*)^q\mathscr{F}$ and let $\tilde{\mathscr{H}} : L^2(D_1) \to L^2(D_2)$ be a HS operator with finite rank $\geq k$ such that $\|\mathscr{H} - \tilde{\mathscr{H}}\| \leq C\sigma_{k+1}^{2q+1}$ for some $C > 0$. Let $\mathcal{U}_k$ be the dominant left k-singular subspace of $\mathscr{H}$, and let $\tilde{\mathcal{U}}_k$ be a dominant left k-singular subspace of $\tilde{\mathscr{H}}$. If $C\delta_q < 1$, then

$$\|\sin\mathbf{\Theta}(\mathcal{U}_k, \tilde{\mathcal{U}}_k)\| \leq \frac{C\delta_q}{1 - C\delta_q}. \tag{45}$$

Proof Let $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots$ be the singular values of $\tilde{\mathscr{H}}$; the singular values of $\mathscr{H}$ are given by $\sigma_1^{2q+1} \geq \sigma_2^{2q+1} \geq \cdots$. By Weyl's theorem, we have $|\tilde{\sigma}_k - \sigma_k^{2q+1}| \leq C\sigma_{k+1}^{2q+1}$, whereas the assumption (44) implies $\sigma_k^{2q+1} - \sigma_{k+1}^{2q+1} = \delta_q^{-1}\sigma_{k+1}^{2q+1}$. It follows that

$$\tilde{\sigma}_k - \sigma_{k+1}^{2q+1} \geq \sigma_k^{2q+1} - |\tilde{\sigma}_k - \sigma_k^{2q+1}| - \sigma_{k+1}^{2q+1} \geq (\delta_q^{-1} - C)\sigma_{k+1}^{2q+1},$$

which in conjunction with Wedin's theorem (Stewart and Sun, 1990, Thm. V.4.4) yields the desired bound. ■

Combining Lemma 13 with rSVD means we can approximate an operator's singular values at the extra cost of computing $\tilde{\mathbf{U}}_k\mathscr{F}$, which takes an additional k operator-function products. This provides the desired proof.

Proof (Theorem 5) Let $\mathbf{U}_k$ be the $D_1 \times k$ quasimatrix whose columns are dominant left k-singular functions of $\mathscr{H}$. A corollary of the CS decomposition (Stewart and Sun, 1990, Exc. I.5.6) along with Theorem 1 and Lemma 13 yields, with high probability,

$$\|\mathbf{U}_k - \tilde{\mathbf{U}}_k\| \leq 2\|\sin \mathbf{\Theta}(\mathcal{U}_k, \tilde{\mathcal{U}_k})\| \leq \frac{2\delta_q A_{k,p}(s,t)}{1-\delta_q A_{k,p}(s,t)}.$$

Hence, we find that

$$\|\mathbf{U}_k^*\mathscr{F} - \tilde{\mathbf{U}}_k^*\mathscr{F}\| \leq \frac{2\delta_q A_{k,p}(s,t)}{1-\delta_q A_{k,p}(s,t)}\|\mathscr{F}\|.$$

Notice that $\mathbf{U}_k^*\mathscr{F}$ has the same dominant k singular values as $\mathscr{F}$, so Weyl's theorem completes the proof. ■

References

T. G. Anderson, O. P. Bruno, and M. Lyon. High-order, dispersionless "fast-hybrid" wave equation solver. Part I: $O(1)$ sampling cost via incident-field windowing and recentering. *SIAM J. Sci. Comput.*, 42(2):A1348–A1379, 2020.

R. Arora. A deep learning framework for solving hyperbolic partial differential equations: Part I. Preprint arXiv:2307.04121, 2023.

J. W. Banks and W. D. Henshaw. Upwind schemes for the wave equation in second-order form. *J. Comput. Phys.*, 231(17):5854–5889, 2012.

M. Bebendorf and W. Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numer. Math.*, 95:1–28, 2003.

J. Berman and B. Peherstorfer. Randomized sparse Neural Galerkin schemes for solving evolution equations with deep networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 4097–4114, 2023.

J. Berman and B. Peherstorfer. CoLoRA: Continuous low-rank adaptation for reduced implicit neural modeling of parameterized partial differential equations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 3565–3583, 2024.

K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

V. I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs.* American Mathematical Society, 1998.

N. Boullé and A. Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Found. Comput. Math.*, 23:709–739, 2023.

N. Boullé, C. J. Earls, and A. Townsend. Data-driven discovery of Green's functions with human-understandable deep learning. *Sci. Rep.*, 12:1–9, 2022a.

N. Boullé, S. Kim, T. Shi, and A. Townsend. Learning Green's functions associated with time-dependent partial differential equations. *J. Mach. Learn. Res.*, 23(1):1–34, 2022b.

N. Boullé, D. Halikias, and A. Townsend. Elliptic PDE learning is provably data-efficient. *Proc. Natl. Acad. Sci. USA*, 120(39):e2303904120, 2023.

N. Boullé, D. Halikias, S. E. Otto, and A. Townsend. Operator learning without the adjoint. *J. Mach. Learn. Res.*, 25(364):1–54, 2024.

J. Bruna, B. Peherstorfer, and E. Vanden-Eijnden. Neural Galerkin schemes with active learning for high-dimensional evolution equations. *J. Comput. Phys.*, 496:112588, 2024.

S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, 113 (15):3932–3937, 2016.

C.-T. Chen and G. X. Gu. Learning hidden elasticity with deep neural networks. *Proc. Natl. Acad. Sci. USA*, 118(31):e2102721118, 2021.

K. Chen, C. Wang, and H. Yang. Deep operator learning lessens the curse of dimensionality for PDEs. In *Transactions on Machine Learning Research*, 2024.

Z. Chen and J. Dongarra. Condition numbers of Gaussian random matrices. *SIAM J. Matrix Anal. Appl.*, 26(3):1389–1404, 2005.

M. Cicognani. Strictly hyperbolic equations with non regular coefficients with respect to time. *Ann. Univ. Ferrara*, 45(1):45–58, 1999.

M. Cicognani and D. Lorenz. Strictly hyperbolic equations with coefficients low-regular in time and smooth in space. *J. Pseudo-Differ. Oper. Appl.*, 9(3):643–675, 2018.

R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 2. Interscience Publishers, 1st English edition, 1962.

M. V. de Hoop, N. B. Kovachki, N. H. Nelsen, and A. M. Stuart. Convergence rates for learning linear operators from noisy data. *SIAM/ASA J. Uncertain. Quantif.*, 11(2): 480–513, 2023.

T. A. Driscoll, N. Hale, and L. N. Trefethen, editors. *Chebfun Guide.* Pafnuty Publications, 2014.

A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010.

Y. Fan and L. Ying. Solving electrical impedance tomography with deep learning. *J. Comput. Phys.*, 404:109119, 2020.

Y. Fan, L. Lin, L. Ying, and L. Zepeda-Núñez. A multiscale neural network based on hierarchical matrices. *Multiscale Model. Simul.*, 17(4):1189–1213, 2019.

H. Federer. Curvature measures. *Trans. Am. Math. Soc.*, 93(3):418–491, 1959.

J. Feliu-Faba, Y. Fan, and L. Ying. Meta-learning pseudo-differential operators with deep neural networks. *J. Comput. Phys.*, 408:109309, 2020.

C. R. Gin, D. E. Shea, S. L. Brunton, and J. N. Kutz. DeepGreen: Deep learning of Green's functions for nonlinear boundary value problems. *Sci. Rep.*, 11:1–14, 2021.

Y. Gordon. Some inequalities for Gaussian processes and applications. *Isr. J. Math.*, 50: 265–289, 1985.

Y. Gordon. Gaussian processes and almost spherical sections of convex bodies. *Ann. Probab.*, 16(1):180–188, 1988.

A. Gray. *Tubes*, volume 221 of *Progress in Mathematics*. Birkhäuser, 2nd edition, 2003.

P. Günther. Huygens' principle and Hadamard's conjecture. *Math. Intell.*, 13(2):56–63, 1991.

Y. Guo, X. Cao, B. Liu, and M. Gao. Solving partial differential equations using deep learning and physical constraints. *Appl. Sci.*, 10(17):5917, 2020.

D. Halikias and A. Townsend. Structured matrix recovery from matrix-vector products. *Numer. Linear Algebra Appl.*, 31(1):e2531, 2024.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53 (2):217–288, 2011.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

A. J. Huang and S. Agarwal. On the limitations of physics-informed deep learning: Illustrations using first-order hyperbolic conservation law-based traffic flow models. *IEEE Open J. Intell. Transp. Syst.*, 4:279–293, 2023.

D. Z. Huang, N. H. Nelsen, and M. Trautner. An operator learning perspective on parameter-to-observable maps. *Found. Data Sci.*, 7(1):163–225, 2025.

D. Hug, G. Last, and W. Weil. A local Steiner-type formula for general closed sets and applications. *Math. Z.*, 246:237–272, 2004.

A. E. Hurd and D. H. Sattinger. Questions of existence and uniqueness for hyperbolic equations with discontinuous coefficients. *Trans. Am. Math. Soc.*, 132(1):159–174, 1968.

A. T. James. Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Stat.*, 35(2):475–501, 1964.

A. T. James. Calculation of zonal polynomial coefficients by use of the Laplace–Beltrami operator. *Ann. Math. Stat.*, 39(5):1711–1718, 1968.

E. Kaiser, J. N. Kutz, and S. L. Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proc. R. Soc. Lond. A*, 474(2219): 20180335, 2018.

G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nat. Rev. Phys.*, 3(6):422–440, 2021.

L. K. Kates. *Zonal polynomials*. PhD thesis, Princeton University, 1981.

Y. Khoo and L. Ying. SwitchNet: a neural network model for forward and inverse scattering problems. *SIAM J. Sci. Comput.*, 41(5):A3182–A3201, 2019.

D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer. Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. USA*, 118(21): e2101784118, 2021.

N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *J. Mach. Learn. Res.*, 24:1–97, 2023.

A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, and M. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 26548–26560, 2021.

J. N. Kutz. Deep learning in fluid dynamics. *J. Fluid Mech.*, 814:1–4, 2017.

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

S. Lanthaler, S. Mishra, and G. E. Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Trans. Math. Appl.*, 6(1):tnac001, 2022.

P. Laurent, G. Legendre, and J. Salomon. On the method of reflections. *Numer. Math.*, 148(2):449–493, 2021.

P. D. Lax. *Hyperbolic Partial Differential Equations*, volume 14 of *Courant Lecture Notes in Mathematics*. American Mathematical Society, 2006.

M. E. Lerner. Qualitative properties of the Riemann function [in Russian]. *Differ. Uravn.*, 27(12):2106–2120, 1991.

J. Levitt and P.-G. Martinsson. Randomized compression of rank-structured matrices accelerated with graph coloring. *J. Comput. Appl. Math.*, 451:116044, 2024.

M. Li, L. Demanet, and L. Zepeda-Núñez. Wide-band butterfly network: stable and efficient inversion via multi-frequency neural networks. *Multiscale Model. Simul.*, 20(4):1191–1227, 2022.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Graph kernel network for partial differential equations. In *International Conference on Learning Representations Workshop on Integration of Deep Neural Models and Differential Equations*, 2020a.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, A. Stuart, K. Bhattacharya, and A. Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6755–6766, 2020b.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

L. Lin, J. Lu, and L. Ying. Fast construction of hierarchical matrix representation from matrix-vector multiplication. *J. Comput. Phys.*, 230(10):4071–4087, 2011.

Y. Liu, X. Xing, H. Guo, E. Michielssen, P. Ghysels, and X. S. Li. Butterfly factorization via randomized matrix-vector multiplications. *SIAM J. Sci. Comput.*, 43(2):A883–A907, 2021.

Y. Liu, J. Song, R. Burridge, and J. Qian. A fast butterfly-compressed Hadamard–Babich integrator for high-frequency Helmholtz equations in inhomogeneous media with arbitrary sources. *Multiscale Model. Simul.*, 21(1):269–308, 2023.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.*, 3:218–229, 2021a.

L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis. DeepXDE: A deep learning library for solving differential equations. *SIAM Rev.*, 63(1):208–228, 2021b.

A. G. Mackie. Green's functions and Riemann's method. *Proc. Edinburgh Math. Soc.*, 14 (4):293–302, 1965.

A. Mandelbaum. Linear estimators and measurable linear transformations on a Hilbert space. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 65(3):385–397, 1984.

S. Massei, L. Robol, and D. Kressner. Hierarchical adaptive low-rank format with applications to discretized partial differential equations. *Numer. Linear Algebra Appl.*, 29(6): e2448, 2022.

M. Meier and Y. Nakatsukasa. Fast randomized numerical rank estimation for numerically low-rank matrices. *Linear Algebra Appl.*, 686:1–32, 2024.

R. Molinaro, Y. Yang, B. Engquist, and S. Mishra. Neural inverse operators for solving PDE inverse problems. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 25105–25139, 2023.

R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, 1982.

NIST DLMF. *NIST Digital Library of Mathematical Functions.* Release 1.1.12 of 2023-12-15. URL `https://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

S. E. Otto, A. Padovan, and C. W. Rowley. Model reduction for nonlinear systems by balanced truncation of state and gradient covariance. *SIAM J. Sci. Comput.*, 45(5): A2325–A2355, 2023.

E. Qian, B. Kramer, B. Peherstorfer, and K. Willcox. Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D*, 406:132401, 2020.

M. Raissi, A. Yazdani, and G. E. Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

M. Reissig. Hyperbolic equations with non-Lipschitz coefficients. *Rend. Semin. Mat. Univ. Politec. Torino*, 61(2):135–181, 2003.

R. Rodriguez-Torrado, P. Ruiz, L. Cueto-Felgueroso, M. C. Green, T. Friesen, S. Matringe, and J. Togelius. Physics-informed attention-based neural network for hyperbolic partial differential equations: Application to the Buckley–Leverett problem. *Sci. Rep.*, 12:7557, 2022.

V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM J. Matrix Anal. Appl.*, 31(3):1100–1124, 2009.

S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Sci. Adv.*, 3(4):e1602614, 2017.

F. Schäfer and H. Owhadi. Sparse recovery of elliptic solvers from matrix-vector products. *SIAM J. Sci. Comput.*, 46(2):A998–A1025, 2024.

F. Schäfer, T. J. Sullivan, and H. Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Model. Simul.*, 19:688–730, 2021.

T. Shi and A. Townsend. On the compressibility of tensors. *SIAM J. Matrix Anal. Appl.*, 42(1):275–298, 2021.

H. F. Smith. A parametrix construction for wave equations with $C^{1,1}$ coefficients. *Ann. Inst. Fourier*, 48(3):797–835, 1998.

G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. Mahoney, and A. Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *Advances in Neural Information Processing Systems*, volume 37, pages 71242–71262, 2023.

B. T. Thodi, S. V. R. Ambadipudi, and S. E. Jabari. Learning-based solutions to nonlinear hyperbolic PDEs: Empirical insights on generalization errors. In *Advances in Neural Information Processing Systems Workshop on Machine Learning and the Physical Sciences*, 2022.

A. Townsend. *Computing with functions in two dimensions.* PhD thesis, University of Oxford, 2014.

A. Townsend and L. N. Trefethen. Continuous analogues of matrix factorizations. *Proc. R. Soc. Lond. A*, 471(2173):20140585, 2015.

L. N. Trefethen. Householder triangularization of a quasimatrix. *IMA J. Numer. Anal.*, 30 (4):887–897, 2010.

L. N. Trefethen. *Approximation Theory and Approximation Practice.* SIAM, extended edition, 2019.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, 2018.

Z. Y. Wan, L. Zepeda-Núñez, A. Boral, and F. Sha. Evolve smoothly, fit consistently: Learning smooth latent dynamics for advection-dominated systems. In *International Conference on Learning Representations*, 2023.

H. Wang. New error bounds for Legendre approximations of differentiable functions. *J. Fourier Anal. Appl.*, 29(42), 2023.

S. Wang, H. Wang, and P. Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Sci. Adv.*, 7:eabi8605, 2021.

L. Zepeda-Núñez and L. Demanet. The method of polarized traces for the 2D Helmholtz equation. *J. Comput. Phys.*, 308:347–388, 2016.

S. Zhang and G. Lin. Robust data-driven recovery of governing physical laws with error bars. *Proc. R. Soc. Lond. A*, 474(2217):20180305, 2018.